NOTE TO USERS

Discrete and statistical approaches to genetics

Trevor Cormac Vincent Bruen

Doctor of Philosophy

Computer Science

McGill University

Montreal, Quebec

2006–10–13

A thesis submitted to McGill University in partial fulfilment of the requirements of the degree of Doctor of Philosophy

Your file Votre référence
ISBN: 978-0-494-32159-1
Our file Notre référence
ISBN: 978-0-494-32159-1

Dedication

To my family.

An aged man is but a paltry thing,

A tattered coat upon a stick, unless

Soul clap its hands and sing, and louder sing

For every tatter in its mortal dress,

Nor is there singing school but studying

Monuments of its own magnificence;

And therefore I have sailed the seas and come

To the holy city of Byzantium.

- W. B. Yeats

Acknowledgements

A number of people have helped make this thesis possible. Firstly, I'd like to thank my thesis advisor David Bryant for his guidance and exposing me to a number of interesting areas of research. Next, I'd like to thank Rachel Bevan for a number of helpful discussions and support. I'd also like to thank Robin and Audrey for support and their help with translating the abstract, as well as the rest of my family and friends for support. Finally, I'd also like to thank all of the people at the McGill Centre for Bioinformatics (past and present), including the members of my advisory committee Mathieu Blanchette and Mike Hallett for providing a warm and stimulating environment.

Preface

This thesis was written according the McGill University requirements as found on the Faculty of Graduate and Post–doctoral Studies website (www.mcgill.ca/gps). I have chosen to write a manuscript based thesis. According to McGill University guidelines, 'as an alternative to the traditional thesis format, the thesis can consist of a collection of papers of which the student is an author or co-author'. These papers can consist of 'the text of one or more papers submitted, or to be submitted, for publication, or the clearly-duplicated text (not the reprints) of one or more published papers.' The manuscripts included in this thesis are as follows. Authorship by the candidate is high–lighted in bold:

Contributions of Authors

Chapter 2: **T. C. Bruen**. and D. Bryant (2006) A subdivision approach to maximum parsimony *In press, Annals of combinatorics.*

The candidate developed and refined an approach initially suggested by Dr. David Bryant. The candidate wrote the paper and the proofs, with some corrections suggested by Dr. David Bryant.

Chapter 3: **T.C. Bruen** and D. Bryant (2006) Maximum parsimony is a consensus method *Submitted, Systematic biology.*

The candidate conjectured and proved the results and wrote the paper based on an area of investigation suggested by Dr. David Bryant who conjectured the two character result as well. Dr. David Bryant also provided helpful comments on the manuscript.

Chapter 4: **T.C. Bruen**, H. Philippe and D. Bryant (2006) A simple and robust statistical test for detecting recombination. *Genetics.* 172:2665-2681.

The candidate proposed the statistic, derived the mean and variance of the statistic, collected the data, proposed and performed the experiments and wrote the paper. Dr. David Bryant suggested the initial idea of using two character parsimony in the context of recombination and made various other suggestions. Dr. Hervé Philippe made various suggestions, including a large simulation study. Both Dr. D. Bryant and Dr. H. Philippe made a number of suggestions on the manuscript.

Chapter 5: **T.C. Bruen** and M. Poss (2006) Recombination shapes the evolution of Feline Immunodeficiency Virus (FIV) in a wild population of cougars. *To be submitted, Journal of Virology.*

The candidate designed and performed the analysis and wrote the paper based on a data set given by Dr. Mary Poss, who also gave a general context for the work. Dr. Mary Poss provided helpful comments on the manuscript.

Abstract

This thesis presents a number of major innovations in related but different areas of research. The contributions range along a continuum from mathematical phylogenetics, to development of statistical methodology for detecting recombination and finally to the application of statistical techniques to understand Feline Immunodeficiency Virus (FIV) an important pathogen. An underlying theme is the application of combinatorial and statistical ideas to problems in evolutionary biology and genetics.

Chapter 2 and Chapter 3 give a number of results relevant to mathematical phylogenetics, in particular maximum parsimony. Chapter 2 presents a new formulation of maximum parsimony in terms of character subdivision, providing a direct link with the character compatibility problem, also known as the perfect phylogeny problem. Specialization of this result to two characters gives a simple formula based on the intersection graph for calculating the parsimony score for a pair of characters. Chapter 3 further explores maximum parsimony. In particular, it is shown that a maximum parsimony tree for a sequence of characters minimizes a subtree-prune and regraft (SPR) distance to the sets of trees on which each character is convex. Similar connections are also drawn between the Robinson-Foulds distance and a new variant of Dollo parsimony.

Chapter 4 presents an application of the work in Chapters 2 and 3 to develop a statistical test for detecting recombination. An extensive

coalescent based simulation study shows that this new test is both robust and powerful in a variety of different circumstances compared to a number of current methods. In fact, a simple model of mutation rate correlation is shown to mislead a number of competing tests, causing recombination to be falsely inferred. Analysis of empirical data sets confirm that the new test is one of the best approaches to distinguish recurrent mutation from recombination.

Finally, Chapter 5 uses the test developed in Chapter 4 to localize recombinant breakpoints in 14 genomic strains of FIV taken from a wild population of cougars. Based on the technique, three recombinant strains of FIV are identified. Previous studies have focused on the epidemiology and population structure of the virus and this study shows that recombination has also played an important role in the evolution of FIV.

Abrégé

Cette thèse présente des innovations majeures dans des domaines de recherches reliés mais différents. Ces contributions s'étendent sur les mathématiques phylogénétiques, au développement de méthodes statistiques pour détecter la recombinaison génétique et sur l'application des techniques statistiques pour comprendre le virus d'immunodéficience féline (VIF) qui est un pathogène important. Un thème sous-jacent est l'application des idées combinatoires et statistiques à des questions de l'évolution de la biologie et de la génétique.

Les chapitres 2 et 3 présentent des résultats pertinents pour les mathématiques phylogénétiques, en particulier vis-à-vis de la parcimonie maximale. Le chapitre 2 montre une nouvelle formulation de la parcimonie maximale pour la sous-division du caractère. Ceci fournit un lien direct avec le problème de la compatibilité du caractère, également connu sous le nom du parfait problème en phylogénie. La spécialisation pour deux caractères résulte en une formule simple basée sur le graphique d'intersection afin de déterminer l'indice de pertinence de la parcimonie pour la pair de caractères. De plus, le chapitre 3 explore la notion de la parcimonie maximale. En particulier, il est démontré qu'un arbre de parcimonie maximale pour une séquence de caractères minimise la distance SPR (pour subtree-prune and regraft) pour les arbres dont chaque caractère est convexe. Des connexions

similaires sont aussi démontrées entre la distance de Robinson-Foulds et une nouvelle variante de la parcimonie de Dollo.

Le chapitre 4 illustre une application des théories abordées aux chapitres 2 et 3 pour développer un test statistique qui puisse déceler la recombinaison génétique. Une étude de simulation coalescente montre que ce nouveau test est robuste et puissant dans différentes circonstances comparé aux méthodes actuelles. En fait, un modèle simple de corrélation du taux de mutation se montre fallacieux vis-à-vis des tests équivalents causant une inférence erronée de la recombinaison. Les analyses de données empiriques permettent de conclure que le nouveau test est l'une des meilleures approches pour distinguer la mutation récurrente de la recombinaison.

Finalement, le chapitre 5 utilise le test développé au chapitre 4 pour localiser des cassures de recombinaison dans 14 souches du VIF présent dans une population sauvage de cougars. Grâce à ce test, trois souches recombinées du FIV ont été identifiées. Les études antérieures se concentraient sur l'épidémologie et la structure de la population du virus alors que la présente étude démontre que la recombinaison génétique joue également un rôle important dans l'évolution du VIF.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1 General background and motivation

Genetics and evolutionary biology are two interrelated disciplines. Genetics can be defined as 'a branch of biology that deals with the heredity and variation of organisms' [24]. Evolution on the other hand can be defined as 'the gradual alteration of an organism or one of its components as a result of genetic changes that are passed from parent to offspring' [37]. In contemporary biology, evolution is often considered the study of macroscopic changes at the species level whereas genetics focuses on the study of organisms within the species level. Of course, the two disciplines are very related and there is a great deal of interplay between them. The study of evolution and genetics is a major focus in biology and provides a general motivation for this work.

The modern theory of evolution was first espoused by Charles Darwin [11]. Darwin suggested that the variety of life forms present did not arise spontaneously but rather had arisen through shared common ancestry [11]. Much like a pedigree can be used to represent family history, a phylogeny can be used to represent relationships between species (Figure 1–1). Indeed,

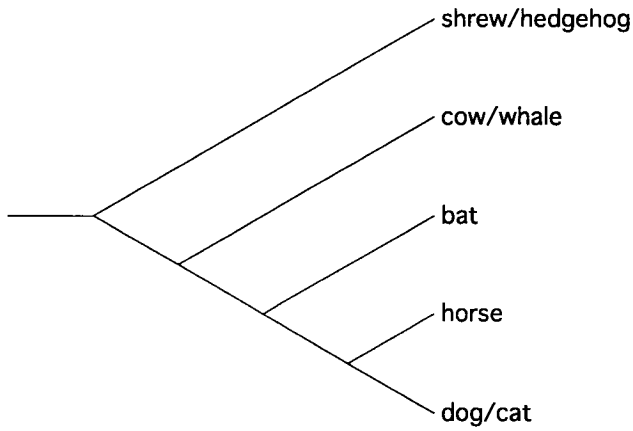


FIGURE 1–1: A phylogeny depicting a controversial part of the mammalian tree [30]. The leaves are labeled with existing species with the branching order indicating relatedness between species.

a phylogeny can be described either as an unrooted or rooted tree, with or without branch lengths, whose leaves are labeled by species (Figure 1–1). This simple idea of a phylogeny had a profound influence in biology and phylogenies are currently used in a broad variety contexts including viral evolution, for instance (see Chapter 5). Phylogenies can also be generalized into networks, the study and application of which is an emerging discipline in phylogenetics [22]. However, the direct use of phylogenies has made an important contribution in bioinformatics ranging from gene finding [31] to regulatory element discovery [4] .

Although phylogenies are useful in their own right, ideas developed in the evolutionary biology context have also been important in genetics. For instance, the idea of (two) character compatibility (defined below) although originating in the evolutionary biology context [27] has had an influence in

the study of recombination as well [21, 29]. Indeed the generalization of this idea plays an important role in our recently developed test for recombination [7].

The development of computational, mathematical and statistical ideas that have relevance to biology is a timely one. The rapid growth in the amount of molecular DNA sequences shows the opportunity and relevance of developing quantitative ideas that can be used to understand biology in general, and evolution and genetics in particular. Ideas about evolutionary history, recombination and selection for instance, require complex mathematical and probabilistic approaches to fully understand them. The development of these computational, mathematical and statistical ideas provides the specific motivation for this work.

1.1.1 Main results of thesis

This thesis presents a number of major innovations in related but different areas of research. The contributions range along a continuum from mathematical phylogenetics, to development of statistical methodology for detecting recombination and finally to the application of statistical techniques to understand (FIV) an important pathogen. An underlying theme is the application of combinatorial and statistical ideas to problems in evolutionary biology and genetics.

In the first two Chapters, the contribution to the mathematical and computational theory of phylogenetics are discussed. In Chapter 2, a new formulation of maximum parsimony (informally taken as the tree that

minimizes the total number of mutations) is given which directly relates it to character compatibility via the intersection graph. Specialization of this result to two characters give a practical formula and algorithm to calculate the parsimony score between a pair of characters. In Chapter 3, a new paradigm of maximum parsimony is presented. In particular, a maximum parsimony tree is shown to have intimate connections with a metric space on the set of trees. The metric that standard parsimony is shown to minimize is the SPR (subtree prune and regraft) distance, although connections are drawn between Dollo parsimony and the Robinson-Foulds distance as well. A specialization to two characters shows that there are connections between the SPR and the parsimony score for a pair of characters. The work in Chapter 3 is relevant to supertree construction and potentially exploration of tree space [20].

The next Chapter, Chapter 4 uses the parsimony score for pairs of characters to develop a new statistic, Φ_w to test for recombination. The mean and variance for the statistic are calculated analytically. Extensive validation shows that testing for recombination based on the statistic is more powerful than previous approaches in a number of circumstances and that false inference of recombination is a serious issue for a number of other approaches. A number of empirical data sets are analyzed, including mitochondrial DNA and the results show that the Φ_w statistic can be used to describe precisely estimate whether or not recombination is present.

Chapter 5 uses the Φ_w statistic to thoroughly examine a single data set, a multiple alignment consisting of strains of FIV. By testing small overlapping regions for recombination, the location of likely breakpoints was found. Further analysis confirmed that the regions did indeed contain likely breakpoints and three recombinant strains were discovered. The results have important implications in the evolution of FIV, both in terms of population structure and epidemiology [2, 3].

1.2 Discrete mathematics in evolutionary biology

Chapters 2 and 3 make a number of contributions to the mathematical understanding of phylogenetics. In this section, the general context of mathematical phylogenetics are presented as well as the notation and main results, so that both Chapters 2 and 3 are understandable.

1.2.1 General overview and context

A number of the original approaches for reconstructing phylogenies used discrete ideas to find optimal trees. Two of the most well-known of these types of approaches are termed character compatibility [27] and maximum parsimony [10]. Both approaches can be used either for molecular (e.g. DNA) or fossil (i.e. morphological) data. Character compatibility (sometimes known as the perfect phylogeny problem) and maximum parsimony have a number of connections with discrete mathematics (see references within [34]). Indeed, the fact that both character compatibility and maximum parsimony have direct connections with graph theory makes

it possible to characterize their computational complexity, for instance [18, 35, 5]. This gives us better insight into the precise difficulty of inferring phylogenies.

In statistical terms, a phylogeny can be viewed as a parameter to be estimated from data. Maximum parsimony and character compatibility are criteria for estimating phylogenies. Given a set of data, each possible phylogeny describing the data can be evaluated under both compatibility and parsimony. The phylogeny that 'maximizes' or 'best fits' the criterion is said to be selected by the criterion. However, no explicit assumptions are made regarding the process of data generation for maximum parsimony and compatibility and thus both criteria may be viewed as 'non-parametric' estimators of the phylogeny.

In a well-known paper, Felsenstein showed that under a simple model of mutation for molecular data, maximum parsimony and compatibility will be in certain cases inconsistent estimators of a phylogeny [14]. That is, in certain cases both maximum parsimony and compatibility will converge in probability to the wrong phylogeny. Felsenstein then proposed estimating a phylogeny which maximized the probability of the data, the maximum likelihood phylogeny [15]. However in a number of cases (such as morphological data or rare insertions in molecular data), there are no good models to explain the data and thus maximum parsimony remains a good option. Moreover, maximum parsimony and character compatibility have found new uses in the supertree context (see Chapter 3 and the references therein).

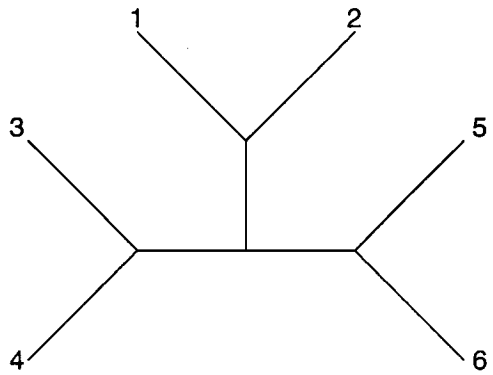


FIGURE 1–2: A binary phylogenetic X–tree, where the leaves are labeled with $X = \{1, 2, 3, 4, 5, 6\}$.

Tuffley and Steel showed that under a very general model (where every site has its own substitution rate) that the maximum likelihood and maximum parsimony phylogeny will be equivalent [36]. Thus, research in this area is not only interesting but timely as well.

1.2.2 Definitions

We follow the notation given by a recent book in the field by Semple and Steel [34]. Let X be a taxon set (informally a set of species). An *X-tree* $\mathcal{T}$ is defined as an ordered pair (T, ϕ) where T is an unrooted tree (an acylic graph) with vertex set V and $\phi : X \rightarrow V$ is a function so that for every vertex $v \in V$ of degree one, $v \in \phi(X)$. A *phylogenetic* X-tree is an X-tree with the property that ϕ is a bijection between X and the leaf set (vertices of degree one) of V. A *binary phylogenetic* X is a phylogenetic X–tree with the property that every internal vertex has degree three (see Figure 1–2).

A *character* χ is defined as a function from a taxon set X to a set C of states, i.e. $\chi : X \rightarrow C$. The number of states of χ (cardinality of the image of χ) is denoted by $|\chi|$; if $|\chi| = 2$ then χ is said to be a binary character. Let $\pi(\chi)$ denote the partition of X induced by $\{\chi^{-1}(\alpha) : \alpha \in C\}$. Each equivalence class of $\pi(\chi)$ is referred to as *block* of χ, with the number of blocks equal to $|\chi|$.

Splits

A *binary character* $\chi : X \rightarrow C$ on X, where $|\chi| = 2$, induces a bipartition of the taxa set. A bipartition of X is termed a *split*. A partition of the species set X into two equivalence classes A and B is denoted by $A|B$. Any edge of an X-tree $\mathcal{T}$ induces a split of X and let $\Sigma(\mathcal{T})$ denote the entire set of splits induced by the edge set of $\mathcal{T}$. The 'Splits Equivalence Theorem' (see below) shows that we may identify an X–tree by its splits.

An *extension* of character χ to a X-tree $\mathcal{T} = (T, \phi)$ is a function $\bar{\chi} : V \rightarrow C$ such that $\bar{\chi} \circ \phi = \chi$. The *change set* of $\bar{\chi}$ on $\mathcal{T}$ is equal to the set $\mathrm{Ch}(\bar{\chi}, \mathcal{T}) = \{\{u, v\} \in E(T) : \bar{\chi}(u) \neq \bar{\chi}(v)\}$. The cardinality of $\mathrm{Ch}(\bar{\chi}, \mathcal{T})$ on $\mathcal{T}$ is denoted as $\mathrm{ch}(\bar{\chi}, \mathcal{T})$. A character χ is said to be *convex* on a X-tree $\mathcal{T}$ if there is an extension $\bar{\chi}$ of χ to $\mathcal{T}$ such that for each $\alpha \in C$ the subgraph induced by $\{v \in V : \bar{\chi}(v) = \alpha\}$ is connected.

1.2.3 Character compatibility

A sequence of characters $\mathcal{C} = (\chi_1, ..., \chi_n)$ are said to *compatible* if there exists an X–tree on which every character is convex. The criterion was

proposed for selecting optimal evolutionary trees implicitly by Le Quesne [27].

The intersection graph int($\mathcal{C}$) for the sequence of characters is defined as a graph with vertex set equal to the ordered pairs:

$$\bigcup_{\chi_i \in \mathcal{C}} \{(\chi_i, A) : A \in \pi(\chi_i)\}.$$

There is an edge between any two vertices whenever the intersection of their second coordinates is non-empty. Note that in particular that any two vertices that share the same character (first coordinate) have no edges between them. A graph is *chordal* if every cycle of at least four vertices contains an edge (a chord) between two non-adjacent vertices. A *chordalization* of a graph $G = (V, E)$ is an addition of edges E', so that $E \subseteq E'$ and $G' = (V, E')$ is chordal. A *restricted chordal completion* of the intersection graph is a chordalization such that there are no edges between vertices that share a second coordinate (blocks from the same character).

A fundamental theorem, originally indicated by Buneman [9], relates a restricted chordal completion of the intersection graph with character compatibility. Formal treatment was given by Steel [35].

Theorem 1.1. *[9, 35] A sequence of characters $\mathcal{C} = (\chi_1, ..., \chi_n)$ are compatible if and only if there is a restricted chordal completion to int($\mathcal{C}$).*

The notion of compatibility plays an important role in Chapter 2. Indeed, direct connections are drawn between the intersection graph for compatibility and maximum parsimony in Chapter 2.

1.2.4 Maximum parsimony

A related criteria for selecting an optimal evolutionary tree is termed maximum parsimony. The general criteria was suggested by Edwards and Cavalli-Sforza and the specific version for multi-state characters that is used here was due to Fitch [12, 17].

Let $\chi : X \rightarrow C$ be character and $\mathcal{T}$ an X–tree. Let $\bar{\chi}$ be a *minimal* extension of χ to $\mathcal{T}$ so that $\mathrm{ch}(\bar{\chi}, \mathcal{T})$ is minimized. Then the value of $\mathrm{ch}(\bar{\chi})$ is called the *parsimony score* of χ on $\mathcal{T}$ and is denoted by $l(\chi, \mathcal{T})$. For a sequence of characters $\mathcal{C} = (\chi_1, ..., \chi_n)$, the X–tree that minimizes

$$l(\mathcal{C}, \mathcal{T}) = \sum_{i=1}^{n} l(\chi_i, \mathcal{T})$$

is called the *maximum parsimony tree* and $l(\mathcal{C}, \mathcal{T})$ is termed the *maximum parsimony score.*

Matrix representation with parsimony

Matrix representation with parsimony refers to a method used by biologists to build supertrees [1, 32]. Informally speaking, the idea is to take a collection of input trees, encode the input trees into binary characters and then apply the parsimony criterion to determine a tree that represents all the input trees. Unfortunately, understanding how parsimony makes sense in this context up to this point has been poorly understood. Chapter 3 gives a new interpretation of parsimony that is relevant in this context.

subsectionDollo parsimony Consider a rooted phylogenetic X-tree T (with edges directed away from the root), and a binary character $\chi : X \rightarrow$

$\{0,1\}$. The *Dollo parsimony score* of χ on T, $l(\chi,T)$ is the minimal length of a *Dollo parsimony extension* $\bar{\chi}$ for which: i) there is at most one edge (u,v) in T with $\bar{\chi}(u)=0$ and $\bar{\chi}(v)=1$, and ii) the number of edges (u,v) such that $\bar{\chi}(u)\neq\bar{\chi}(v)$ is minimized (hence 1 is referred to as the derived state).

A Lemma of Huson and Steel [23] allows the DP-score of unrooted trees to be considered. Let $T^{-\rho}$ denote the unrooted phylogenetic X-tree obtained by suppressing the root vertex ρ. The following result allows us to consider unrooted trees rather than rooted trees.

Lemma 1.1. [23] *For a rooted phylogenetic X-tree T and a character $\chi: X\to\{0,1\}$ we have*

$$l_{DP}(\chi,T)=\Delta(\chi,T^{-\rho})$$

Thus, $l_{DP}(\chi,T)$ is independent of the placement of a root.

Then for an *unrooted* phylogenetic X–tree T, we can define the DP score for T as $l_{DP}(\chi,T^{+\rho})$, where $T^{+\rho}$ is an arbitrary rooting of T. By Lemma 1.1, this notion is well-defined.

Homoplasy

In a rooted tree, biologists use the notion of homoplasy as a measure of the number of recurrent or convergent mutations in a tree (Figure 1–3). For an extension $\bar{\chi}$, a mutation is termed *recurrent* if by directing all edges away from the root, there is a path $v_1,...,v_k$ such that $\bar{\chi}(v_1)=\bar{\chi}(v_k)$ and $\bar{\chi}(v_i)\neq\bar{\chi}(v_1)$ when $1<i<k$ [34]. Under the same conditions, a mutation

is termed *convergent* if there are two paths $v_1, ..., v_k$ and $w_1, ..., w_l$ that are vertex disjoint except for $w_1 = v_1$ and where $\bar{\chi}(v_k) = \bar{\chi}(w_l) \neq \bar{\chi}(v_1)$ [34].

To formalize this notion we restate a proposition of Semple and Steel [34].

Lemma 1.2. *[34] Let χ be a character on X and let $\mathcal{T}$ be an X–tree. Then*

$$l(\chi, \mathcal{T}) \geq |\chi| - 1$$

Moreover equality holds if and only if χ is convex on $\mathcal{T}$.

This can be seen informally by noting that every state (all $|\chi|$ states) must be present in the tree and thus there must be at least $|\chi|-1$ edges with changes in the tree. Lemma 1.2 allows us to define the notion of *homoplasy* of character χ on an X-tree $\mathcal{T}$. The homoplasy of χ on $\mathcal{T}$ is defined as:

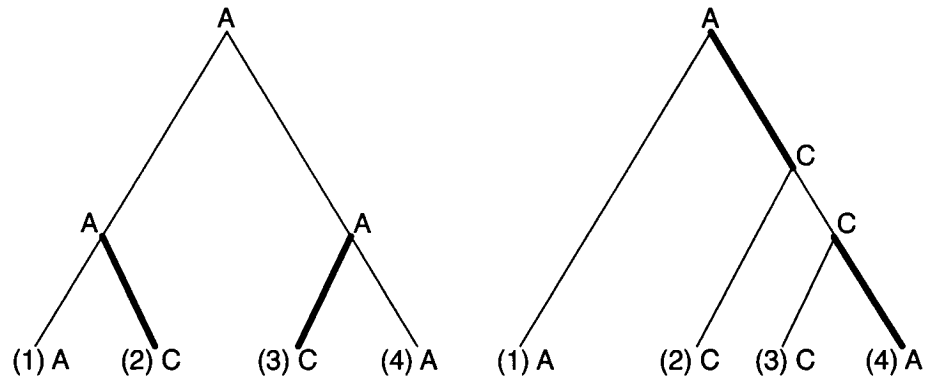
$$h(\chi, \mathcal{T}) = l(\chi, \mathcal{T}) - |\chi| + 1$$

By Lemma 1.2, we have that $h(\chi, \mathcal{T}) \geq 0$.

Note that $l(\chi, \mathcal{T}) = h(\chi, \mathcal{T}) + |\chi| - 1$ and $|\chi| - 1$ is constant over all X–trees $\mathcal{T}$. Hence any maximum parsimony tree $\mathcal{T}$ for a sequence of characters $\mathcal{C} = (\chi_1, ..., \chi_n)$ must equivalently minimize the total homoplasy, i.e.:

$$h(\mathcal{C}, \mathcal{T}) = \sum_{i=1}^{n} h(\chi_i, \mathcal{T})$$

Indeed, the primary motivation of maximum parsimony from a biological perspective is to minimize homoplasy, that is convergent or recurrent mutations. Note that by Lemma 1.2 determining whether a sequence of



(a) Illustration of a convergent mutation (b) Illustration of a recurrent mutation

FIGURE 1–3: A leaf-labeled phylogenetic X–tree, $\mathcal{T}$ with $X = \{1, 2, 3, 4\}$ with $\chi(1) = \chi(4) = A$ and $\chi(2) = \chi(3) = C$ so that $C = \{A, C\}$. A minimal extension of χ is shown with the bold edges indicating the mutations that can be termed convergent and recurrent respectively.

characters $\mathcal{C} = (\chi_1, ..., \chi_n)$ is compatible is equivalent to determining whether there is an X–tree with total homoplasy of 0. The notion of homoplasy plays an important role in Chapter 3.

Two character compatibility

The compatibility of a pair of characters, either binary or multi-state can be determined straightforwardly in linear time [27, 13]. But to determine whether a sequence of characters $\mathcal{C} = (\chi_1, ..., \chi_n)$ are compatible is NP-hard in the general case [35, 6]). However, if all characters are binary it is possible to determine their compatibility in linear time [9, 28].

Indeed, the reason that the compatibility of the entire sequence of binary characters can be determined easily is due to the fact that compatibility for the sequence depends only on the pairwise compatibility

of all members of the set, a fundamental result in the field, known as the 'Splits Equivalence Theorem' due to Buneman [8, 34]. The result is stated in terms of splits that have a very close relationship with binary characters; two splits of X, $A|B$ and $C|D$ are said to be compatible if and only if at least one of $A \cap C$, $A \cap D$, $B \cap C$ or $B \cap D$ is empty (see [34] for further details).

Theorem 1.2. *[8, 34] Let Σ be a collection of splits. Then, there is an X–tree T such that $\Sigma = \Sigma(T)$ if and only if the splits in Σ are pairwise compatible. Moreover, if such an X-tree exists, then, up to isomorphism, T is unique.*

Recall that determining a compatible tree for an arbitrary sequence of characters is equivalent to determining a maximum parsimony tree with no homoplasy. This hold for two characters as well; determining whether two characters are compatible is equivalent to determining whether there is a maximum parsimony tree for both characters with no homoplasy. In Chapter 2, the notion of two character parsimony is explored further resulting in a simple formula and linear time algorithm for the maximum parsimony score for a pair of characters.

1.2.5 Metrics in tree space

Denote the set of all binary phylogenetic X–trees on a taxa set X, by $UB(X)$ where $|X| = n$. It is natural then to consider metrics d defined on $UB(X)$ to compare arbitrary pairs of trees. In other words, we need to

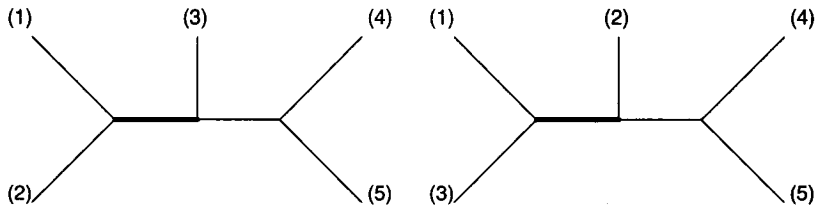


FIGURE 1–4: Two binary phylogenetic X–trees $\mathcal{T}_1$ and $\mathcal{T}_2$ where $X = \{1, 2, 3, 4, 5\}$ and $d_{RF}(\mathcal{T}_1, \mathcal{T}_2) = 2$. Each tree differs by the other tree by one split and the edge representing each of the distinct splits is darkened.

be able to compare two arbitrary binary phylogenetic X–trees. There are several ways to do this.

Robinson-Foulds distance

Given two X-trees $\mathcal{T}_1$ and $\mathcal{T}_2$, define the *Robinson-Foulds distance* (or symmetric difference) between two trees as $d_{RF}(\mathcal{T}_1, \mathcal{T}_2) = |\Sigma(\mathcal{T}_1)\Delta\Sigma(\mathcal{T}_2)|$ (Figure 1–4) [33] (where Δ is the set symmetric difference). By the 'Splits Equivalence Theorem', an X–tree can be identified with its splits, and so the Robinson-Foulds distance can be shown to constitute a metric. Informally speaking, the Robinson-Foulds distance between two X–trees constitutes the cardinality of the set edges that are present in one tree but absent in the other.

Subtree-prune and regraft distance

This metric arises after considering the notion of a *subtree-prune and regraft* (SPR). Consider an X–tree $\mathcal{T} = (T, \phi)$ with $E(T)$ consisting of the edges of T. A subtree-prune and regraft begins by removing an edge $e = \{u, v\} \in E(T)$, leaving two disconnected components, including the subtree containing u. Then an arbitrary edge in the component that

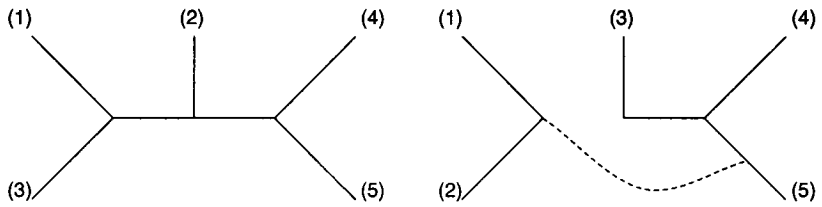


FIGURE 1–5: Two binary phylogenetic X–trees $\mathcal{T}_1$ and $\mathcal{T}_2$ where $X = \{1, 2, 3, 4, 5\}$ and $d_{SPR}(\mathcal{T}_1, \mathcal{T}_2) = 1$. The second tree $\mathcal{T}_2$ differs from the first trees $\mathcal{T}_1$ by one SPR operation, the reconnected edge shown with a dashed line.

contains v is chosen, a vertex v' in the middle of the chosen edge is added and both disconnected components are re-connected by adding an edge $\{u, v'\}$. The resulting tree is said to differ by one SPR from the original tree (Figure 1–5).

Given two X-trees $\mathcal{T}_1 = (T_1, \phi)$ and $\mathcal{T}_2 = (T_2, \phi')$, define the subtree-prune and regraft distance, d_{SPR} as the minimum number of SPRs needed to transform T_1 to T_2. It is easily seen that this conforms to the definition of a metric.

Tree-bisection and reconnection distance

The tree–bisection and reconnection (TBR) metric is essentially identical to the subtree prune and regraft distance, except a more general operation to transform trees is used. Consider an X–tree $\mathcal{T} = (T, \phi)$ with $E(T)$ consisting of the edges of T. A tree-bisection and recombination begins by removing an edge $e = \{u, v\} \in E(T)$, leaving two disconnected components. Then two arbitrary edges are chosen, one in the component that contains u and the other in the component that contains v are chosen.

Two vertices u' and v' are added in the middle of the two chosen edges, respectively. Finally both disconnected components are re-connected by adding an edge $\{u', v'\}$. The resulting tree is said to differ by one TBR from the original tree.

1.2.6 Basic Coalescent with Wright-Fisher Model

In order to understand the evolution through time of a random sample of genes (or loci) from a population, a stochastic process known at the coalescent is commonly used [38]. The evolution of a sample is represented by a genealogy which is similar to a rooted evolutionary tree where the root to tip distance is constant. The idea is to represent individuals by a gene or other genetic markers. The coalescent process gives a description of the genealogy of a random sample of size n that is a subset of a total population of size N. But for the coalescent process to be properly defined, a model for population reproduction must first be assumed. A common model for population reproduction is the Wright-Fisher model [16, 39]. The basic idea of the Wright-Fisher model is that there are discrete non-overlapping generations. Lineages in the current generation are obtained by randomly sampling with replacement from lineages in the previous generation. This corresponds to a notion of random mating.

The coalescent is the limiting process of the Wright-Fisher model as N goes to infinity which parametrically describes the genealogies for a random sample of genes through time [38]. Given a random sample of size 2 in the current generation and looking backwards in time, the two members are

said to coalesce at the first point (in time) at which they share a common ancestor; this is called a coalescence time. For a sample of size n, the values of T_1, ..., T_n are said to represent coalescence times; that is the points at which the distinct number of lineages decrease by one. Kingman showed that the times to coalescence are independent and exponentially distributed as [38, 25, 26]:

$$f_{T_i}(t_i) = \binom{i}{2} e^{\binom{i}{2} t_i}$$

where $t_i \geq 0$, $i = 2, ..., n$. For this process, there is an inherent assumption of no population structure and no fitness difference among mutations.

The genealogy of the random sample is unobserved; instead the random sample along with its mutations are observed. Part of the attraction of the coalescent process is that the genealogical process and mutation process are independent [38]. This is important, since often the genealogy itself is of little interest, in contrast to evolutionary biology. Instead, the interest lies in estimates of parameters such as the population mutation rate θ, which governs the observed amount of mutation. These are obtained roughly speaking by integrating over genealogies (see [19, 38] for further details). The coalescent process can be used to generate random samples from the current generation as well infer parameters such as θ and ρ (see Chapter 4).

References

[1] B. R. Baum. Combining trees as a way of combining datasets for phylogenetic inference, and the desirability of combining gene trees. *Taxon*, 41:3–10, 1992.

[2] R. Biek, A. J. Drummond, and M. Poss. A virus reveals population structure and recent demographic history of its carnivore host. *Science*, 311(5760):538–541, 2006.

[3] R. Biek, A.G. Rodrigo, D. Holley, A. Drummond, C. R. Anderson Jr., H. A. Ross, and M. Poss. Epidemiology, genetic diversity, and evolution of endemic feline immunodeficiency virus in a population of wild cougars. *Journal of Virology*, 77(17):9578–9589, 2003.

[4] M. Blanchette and M. Tompa. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Research*, 12(5):739–748, 2002.

[5] H. Bodlaender, M. Fellows, M. Hallett, T. H. Wareham, and T. Warnow. The hardness of Perfect Phylogeny, Feasible Register Assignment and other problems on thin colored graphs. *Journal of Theoretical Computer Science*, 244:167–188, 2000.

[6] H. L. Bodlaender, M. R. Fellows, and T. J. Warnow. Two strikes against perfect phylogeny. In *Proceedings of the Internation Colloquium*

on Automata, Languages and Programming, volume 623, pages 273–283, Berlin, 1992. Springer-Verlag.

[7] T. C. Bruen, H. Philippe, and D. Bryant. A simple and robust statistical test for detecting recombination. *Genetics*, 172:2665–2681, 2006.

[8] P. Buneman. The recovery of trees from measure of dissimilarity. In D.G. Kendall F.R. Hodson and P. Tautu, editors, *Mathematics in the Archaelogical and Historical Sciences.* Edinburgh University Press, Edinburgh, pp. 387-395, 1971.

[9] Peter Buneman. A characterisation of rigid circuit graphs. *Discrete mathematics*, 9:205–212, 1974.

[10] J. H. Camin and R. R. Sokal. A method for deducing branching sequences in phylogeny. *Evolution*, 19(3):311–326, 1965.

[11] C. Darwin. *On the origin of species by means of natural selection.* John Murray, Albermarle Street, London, U.K., 1859.

[12] A. W. F. Edwards and L. L. Cavalli-Sforza. The reconstruction of evolution. *Annals of Human Genetics*, 27:105–106, 1963.

[13] G. F. Estabrook and F. R. McMorris. When are two qualitative taxonomic characters compatible? *Journal of Mathematical Biology*, 4:195–200, 1977.

[14] J. Felsenstein. Cases in which parsimony and compatibility methods will be positively misleading. *Systematic Zoology*, 27:401–410, 1978.

[15] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Systematic Zoology*, 17:368–376, 1981.

[16] R. A. Fisher. *The Genetical Theory of Natural Selection.* Clarendon Press, Oxford., 1930.

[17] W. M. Fitch. Towards defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, 20(4):406–416, 1971.

[18] L. R. Foulds and R. L. Graham. The Steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics*, 3:43–49, 1982.

[19] J. Hein, M. H. Schierup, and C. Wiuf. *Gene Genealogies, Variation and Evolution.* Oxford University Press, 2005.

[20] D. M. Hillis, T. Heath, and K. St. John. Analysis and visualization of tree space. *Systematic Biology*, 54:471–482, 2005.

[21] R. R. Hudson and N. L. Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111(1):147–64, 1985.

[22] D. H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–267, 2006.

[23] D. H. Huson and M. A. Steel. Phylogenetic trees based on gene content. *Bioinformatics*, 20:2044–9, 2004.

[24] Merriam-Webster Incorporated. *Merriam Webster's Collegiate Dictionary, 10th edition.* Merriam-Webster, Incorporated, Springfield,

Massachusetts, U.S.A., 1994.

[25] J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982.

[26] J. F. C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19:27–43, 1982.

[27] W. J. Le Quesne. A method of selection of characters in numerical taxonomy. *Systematic Zoology*, 18(2):201–205, 1969.

[28] C. A. Meacham. Theoretical and computational considerations of the compatibility of qualitative taxonomic characters. In J. Felsenstein, editor, *Numerical Taxonomy*, volume G1 of *NATO ASI Series*, pages 304–314, Berlin, 1983. Springer-Verlag.

[29] S. R. Myers and R. C. Griffiths. Bounds on the minimum number of recombination events in a sample history. *Genetics*, 163(1):375–94, 2003.

[30] H. Nishihara, M. Hasegawa, and N. Okada. Pegasoferae, an unexpected mammalian clade revealed by tracking anceint retroposon insertions. *Proceedings of the National Academy of Sciences of the United States of America*, 103:9929–9934, 2006.

[31] J. K. Pedersen and J. Hein. Gene finding with a hidden markov model of genome structure and evolution. *Bioinformatics*, 19(2):219–227, 2003.

[32] M. A. Ragan. Phylogenetic inference based on matrix representations of trees. *Molecular Phylogenetics and Evolution*, 1:53–58, 1992.

[33] D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.

[34] C. Semple and M. A. Steel. *Phylogenetics.* Oxford University Press, 2003.

[35] M. A. Steel. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification*, 9:91–116, 1992.

[36] C. Tuffley and M. Steel. Links between maximum likelihood parsimony under a simple model of site substitution. *Bulletin of mathematical biology*, 59:581–607, 1997.

[37] D. Voet, J. G. Voet, and C. W. Pratt. *Fundamentals of Biochemistry.* John Wiley and Sons, Inc., New York, New York, U.S.A., 1999.

[38] J. Wakeley. *Coalescent Theory: An Introduction.* Roberts and Company Publishers, 2004.

[39] S. Wright. Evolution in Mendelian populations. *Genetics*, 67:97–159, 1931.

CHAPTER 2

A subdivision approach to maximum parsimony

2.1 Background

This chapter presents a 'dual formulation' of maximum parsimony which links it directly to maximum compatibility via the intersection graph. A practical result is an analytical formulation of the two character parsimony score which is further explored in Chapter 3 and used in Chapter 4.

2.2 Abstract

Determining an optimal phylogenetic tree using maximum parsimony, also referred to as the Steiner tree problem in phylogenetics, is NP hard. Here we provide a new formulation for this problem that leads to an analytical and linear time solution when the dimensionality (sequence length, or number of characters) is at most two. This new formulation of the problem provides a direct link between the maximum parsimony problem and maximum compatibility problem via the intersection graph. The solution for the 'two character case' has numerous practical applications in phylogenetics, some of which are discussed.

2.3 Introduction

Given a connected graph $G = (V, E)$, an edge weight $w(e) \in \mathbb{Z}_0^+$ for each $e \in E$ and a set of vertices $S \subseteq V$, the *Steiner tree* problem is to find a subtree $T = (V', E')$ of G such that $S \subseteq V'$ and the sum of all the edge weights is minimized [8]. It is well known to be NP complete [12]. A more restricted version of the general problem can be obtained by insisting the edge weights conform to some metric. For instance, consider a fixed alphabet A and the complete graph G on A^N (N is referred to here as the dimension) with edge weights defined as the Hamming distance d on A^N, i.e. $d((a_1, ..., a_n), (a'_1, ..., a'_n))$ is equal to the number of indices i such that $a_i \neq a'_i$. Then the phylogenetic Steiner tree problem (or maximum parsimony problem) is to find a Steiner tree for G whose vertex set includes $S \subseteq A^N$. This problem is also known to be NP complete [7].

For the most part, statistical methods for inferring phylogenies [6, 11] have supplanted maximum parsimony approaches in the construction of phylogenetic trees from conventional sequence data. Nevertheless, maximum parsimony is still widely used to infer evolutionary trees based on morphological characters, to build supertrees, and to perform fast heuristic tree searches.

In this note we make two principal contributions: i) an alternate formulation of the maximum parsimony problem in terms of subdivision; and ii) detailed analysis of the two dimensional case (i.e. $N = 2$) showing that the problem can be solved not only in polynomial time, but actually

linear time. The latter result is proved in two different ways and upper bounds on the maximum parsimony score for two characters are derived. The two character results permit an efficient approach to detecting genetic recombination [2], although there are potentially other applications such as improved lower bounds for parsimony (e.g. [15, 10]). The result on two characters permits the efficient computation of the refined incompatibility score [3, 13] for two characters.

2.4 Notation and Definitions

Further details concerning mathematical phylogenetics and origin of the notation can be found elsewhere [14]. Let X be a set of n species and χ a function (called a *character*) from X to a finite set of states C. The number of states of χ (cardinality of the image of χ) is denoted by $|\chi|$. Let $\pi(\chi)$ denote the partition of X induced by $\{\chi^{-1}(\alpha) : \alpha \in C\}$. Each equivalence class of $\pi(\chi)$ is referred to as *block* of χ, with the number of blocks equal to $|\chi|$. A character χ' *refines* χ if every block of χ' is a subset of some block of χ, which holds if and only if $\chi'(u) = \chi'(v)$ implies $\chi(u) = \chi(v)$ for all $u, v \in X$. Note that the character with only one block is refined by all other characters, while the character with one block for each element in X refines all other characters. A *subdivision* of a character χ is the replacement of one block of the character with two disjoint and non-empty blocks.

An *X–tree* is an ordered pair $\mathcal{T} = (T, \phi)$ consisting of a tree T and a function $\phi : X \to V(T)$ with the property that every vertex of T of degree 1 or 2 is a labeled by ϕ. A *phylogenetic* X–tree is a X–tree with the property

that ϕ induces a bijection between X and the leaves of T . An *extension* of χ to an X–tree $\mathcal{T} = (T, \phi)$, is a function $\bar{\chi}$ from $V(T)$ to C such that the vertices of T are labeled in accordance with χ, i.e. $\bar{\chi} \circ \phi = \chi$.

Consider an extension $\bar{\chi}$ of some character χ to an X–tree $\mathcal{T}$ with underlying tree T. Then define $\mathrm{Ch}(\bar{\chi}, \mathcal{T}) := \{e = \{u, v\} \in E(T) : \bar{\chi}(u) \neq \bar{\chi}(v)\}$ and $\mathrm{ch}(\bar{\chi}, \mathcal{T}) := |\mathrm{Ch}(\bar{\chi}, \mathcal{T})|$. The *parsimony score* of χ on $\mathcal{T}$, $l_{\mathcal{T}}(\chi)$, is defined as the minimum of $\mathrm{ch}(\bar{\chi}, \mathcal{T})$ over all extensions of χ to $\mathcal{T}$. A character χ is *convex* on an X–tree $\mathcal{T}$ if and only if $l_{\mathcal{T}}(\chi) = |\chi| - 1$. For a sequence of k characters $\mathcal{C} = (\chi_1, ..., \chi_k)$ and an X-tree $\mathcal{T}$, the parsimony score of $\mathcal{C}$ on $\mathcal{T}$, $l_{\mathcal{T}}(\mathcal{C})$ is equal to the sum of $l_{\mathcal{T}}(\chi_i)$ for $1 \leq i \leq k$. An X–tree $\mathcal{T}$ that minimizes $l_{\mathcal{T}}(\mathcal{C})$ is said to be a *maximum parsimony tree*, and the minimum value of $l_{\mathcal{T}}(\mathcal{C})$, written as $l(\mathcal{C})$, is said to be the *maximum parsimony score.* A sequence of characters $\mathcal{C}$ are said to be *compatible* if and only if there is some X–tree $\mathcal{T}$ on which every character is convex.

The parsimony score of two characters can be used to calculate $i(\chi_1, \chi_2)$ the *refined incompatibility score* [2], defined as

$$i(\chi_1, \chi_2) = l(\chi_1, \chi_2) - |\chi_1| - |\chi_2| + 2.$$

2.5 Subdivision formulation of parsimony

In this section we reformulate the maximum parsimony criterion in terms of minimal convex refinements, or minimal subdivisions.

Lemma 2.1. *Let $\chi : X \to C$ be a multi–state character and $\mathcal{T}$ a phylogenetic tree. Then there exists a refinement χ' of χ such that $|\chi'| = l_{\mathcal{T}}(\chi) + 1$ and χ' is convex on $\mathcal{T}$.*

Proof. Let $\bar{\chi}$ be a minimal extension of χ to $\mathcal{T}$. Removing the $l_{\mathcal{T}}(\chi)$ edges in $\mathrm{Ch}(\bar{\chi}, \mathcal{T})$ gives $l_{\mathcal{T}}(\chi) + 1$ connected components of T on which $\bar{\chi}$ is constant. As $\bar{\chi}$ is minimal, each component must contain at least one leaf. Define a new character χ' such that the blocks of χ' are in correspondence with the subset of taxa that label the leaves of each connected component. Then χ' has $l_{\mathcal{T}}(\chi) + 1$ blocks, is convex on $\mathcal{T}$ and if $\chi'(x) = \chi'(y)$ then $\chi(x) = \chi(y)$, so χ' refines χ. □

Lemma 2.2. *Let χ be a multi–state character on X and $\mathcal{T}$ a phylogenetic tree. Let χ' be any character that is convex on $\mathcal{T}$ and refines χ. Then $l_{\mathcal{T}}(\chi) \leq |\chi'| - 1$.*

Proof. Let $\bar{\chi'}$ be a minimal extension of χ' to $\mathcal{T}$. Since χ' is convex, removing the edges of $\mathrm{Ch}(\bar{\chi'}, \mathcal{T})$ gives $|\chi'|$ connected components that each contain at least one leaf. Define an extension $\bar{\chi}$ of χ to $\mathcal{T}$ by $\bar{\chi}(v) = \chi(l)$ where v is any vertex and l is any leaf in the component that contains v. This extension is well-defined since $\bar{\chi'}$ and hence χ is constant on the leaves of each component. Since $\bar{\chi}$ is constant on every component we have that $\mathrm{ch}(\bar{\chi}, \mathcal{T}) \leq |\chi'| - 1$ and so $l_{\mathcal{T}}(\chi) \leq |\chi'| - 1$.

□

Theorem 2.1. *Let $\mathcal{C} = (\chi_1, ..., \chi_k)$ be a sequence of k characters on X and let $l(\mathcal{C})$ denote the maximum parsimony score. Let B denote the minimum of $\sum_{i=1}^{k} |\chi_i'|$ over all characters $\chi_1', ..., \chi_k'$ that refine $\chi_1, ..., \chi_k$ respectively and are convex over some tree $\mathcal{T}$. Then $l(\mathcal{C}) = B - k$.*

Proof. Let $\mathcal{T}$ be a maximum parsimony phylogenetic X–tree for $\mathcal{C}$ (note that we may assume $\mathcal{T}$ is a phylogenetic X–tree since such a tree can be readily obtained from a non-phylogenetic X–tree). Then by Lemma 1 there exist refinements $\chi_1', ..., \chi_k'$ of $\chi_1, ..., \chi_k$ that are convex on $\mathcal{T}$ such that $|\chi_i'| = l_{\mathcal{T}}(\chi_i') + 1$. So

$$B \leq l(\mathcal{C}) + k.$$

On the other hand, let $\chi_1', ..., \chi_k'$ be any characters that refine $\chi_1, ..., \chi_k$, are convex on some phylogenetic X–tree $\mathcal{T}$ and satisfy $B = \sum_{i=1}^{k} |\chi_i'|$. Then by Lemma 2 $l_{\mathcal{T}}(\chi_i) \leq |\chi_i'| - 1$ and so

$$l(\mathcal{C}) \leq \sum_{i=1}^{k} l_{\mathcal{T}}(\chi_i) \leq B - k.$$

□

Theorem 2.1 can be reformulated in terms of character subdivisions, noting that each subdivision increases the number of blocks by one, and that if χ' refines χ then χ' can be obtained from χ through a series of subdivisions. Hence we have

Corollary 2.1. *Let $\mathcal{C} = (\chi_1, ..., \chi_k)$ be a sequence of k characters on X. Then the parsimony score is equal to $\sum_i(|\chi_i| - 1)$ plus the number of subdivisions required to transform $\mathcal{C}$ into a sequence of compatible characters.*

In other words, the parsimony score for a sequence of characters equals the minimum number of subdivision required for those characters to have a *perfect phylogeny*, in the technical sense (e.g. [9]).

2.6 Two characters - intersection graph approach

We now turn our attention to the problem of computing parsimony scores for pairs of characters. For this, we draw on connections between characters and intersection graphs [5]. The *intersection graph* for two characters χ_1 and χ_2 has one vertex for every block of χ_1 and χ_2 and an edge between vertices corresponding to blocks that have a non-empty intersection [14]. We denote this graph by $\Gamma(\chi_1, \chi_2)$. Clearly, $\Gamma(\chi_1, \chi_2)$ is bipartite. The theorem we need can be stated as (from [5]):

Theorem 2.2. [5] *Two characters χ_1 and χ_2 on X are compatible if and only if $\Gamma(\chi_1, \chi_2)$ is acyclic.*

Theorem 2.3 can be viewed as a generalisation of Theorem 2.2.

Theorem 2.3. *Let χ_1 and χ_2 be two multi–state characters and $\Gamma(\chi_1, \chi_2) = (V, E)$ the intersection graph for the two characters. Then the maximum parsimony tree for χ_1 and χ_2 has score $|E| + K - 2$, where K is the number of components in $\Gamma(\chi_1, \chi_2)$.*

Proof. Let χ_1' and χ_2' be refinements of χ_1 and χ_2 that are convex on some tree and let $\Gamma(\chi_1', \chi_2') = (V', E')$ be the corresponding intersection graph. Let K' be the number of components of (V', E'). Note that $|V'| \geq |V|$, $|E'| \geq |E|$ and $K' \geq K$ since refining a character cannot decrease any of these quantities. As (V', E') is acyclic we have $|V'| = |E'| + K'$. Hence $|V'| \geq |E| + K$ and, by Theorems 2.1 and 2.2 the maximum parsimony score is at least $|E| + K - 2$.

To show that this minimum can be achieved, it is sufficient to show that if $|E| + K - |V| > 0$, then one of the two characters can be subdivided so that $|V|$ increases by 1 with K and $|E|$ constant. Repeated subdivisions will then achieve the desired minimum.

If $|E| + K - |V| > 0$ then (V, E) contains a cycle. Let $\{w, u\}$ be any edge lying on the cycle, where w corresponds to a block B_1 of χ_1 and u corresponds to a block B_2 of χ_2. As w lies on a cycle of $\Gamma(\chi_1, \chi_2)$ we have that $B_1 - B_2$ is non-empty. Subdivide B_1 into two blocks $B_1 \cap B_2$ and $B_1 - B_2$. The effect on $\Gamma(\chi_1, \chi_2)$ is to replace w by two vertices w_1 and w_2 so that there is an edge $\{w_1, u\}$ and if $\{w, y\}$ is any edge in the old graph, where $y \neq u$, then $\{w_2, y\}$ is an edge in the new graph. The number of edges has not increased. Furthermore, there is a path from u to w_2 along the other edges in the cycle and hence a path from w_1 to w_2. This implies the number of components has not increased either. Therefore we have found a subdivision that increases $|V|$ by 1 but leaves the number of edges and the number of components constant. Repeating this procedure gives a pair of

characters χ'_1 and χ'_2 with $\Gamma(\chi'_1, \chi'_2) = (V', E')$ and K' components where $|E'| + K' - |V'| = 0$ with Γ acyclic. By Theorems 1 and 2 the parsimony score for the pair of characters is then $|V'| - 2$ or $|E| + K - 2$. □

Note that the linear time calculation for the parsimony score follows from the fact that the intersection graph can be constructed in $O(n)$ time and a depth first search to count the number of components in the graph takes $O(n)$ time, where n is the number of taxa (i.e. $|X| = n$). Interestingly, up to this point, the determination of compatibility of two multi-state characters has implicitly been described as a breadth first search for a cycle within an intersection graph [4].

Note that using the framework of Theorem 2.3, the refined incompatibility score for two characters is equal to $|E|+K-|V|$ since $|\chi_1|+|\chi_2| = |V|$. Another result that arises from Theorem 2.3 concerns the upper bound on the maximum parsimony score for two characters.

Corollary 2.2. *Let χ_1 and χ_2 be any two characters on X with $|\chi_1| = r_1$ and $|\chi_2| = r_2$. Then the maximum parsimony score for χ_1 and χ_2 is bounded above by $r_1 r_2 - 1$.*

Proof. By Theorem 2.3 the maximum parsimony score is equal to $|E|+K-2$ where $|E|$ denotes the number of edges and K denotes the number of components in $\Gamma(\chi_1, \chi_2)$. If $K = 1$ it is easily seen that $r_1 r_2 - 1$ is an upper bound since $\Gamma(\chi_1, \chi_2)$ is a bipartite graph with r_1 and r_2 vertices in each part. Note that adding an edge between any two components cannot decrease the parsimony score. Hence, it is sufficient to consider the case of

one component since any upper bound for the many component case is less than or equal to the upper bound for the one component case. □

The upper bound in Corollary 2.2 is tight. Set

$$X = \{x_{ij} : 1 \leq i \leq r_1,\ 1 \leq j \leq r_2\}$$

let χ_1 be the character taking x_{ij} to i and let χ_2 be the character taking x_{ij} to j, for all i, j. Then $\Gamma(\chi_1, \chi_2)$ has one component and $r_1 r_2$ edges, so that the pair of characters has parsimony score $r_1 r_2 - 1$.

2.7 Two characters - spanning tree approach

We now explore the relationship between parsimony trees for two characters and minimum spanning trees, similar to ideas presented in Proposition 5.4.1 of a recent book [14]. A crucial distinction is that Proposition 5.4.1 is stated for a general metric space setting of parsimony (see [14] for details). The following lemma and theorem implicitly assume the discrete metric space setting for parsimony.

Lemma 2.3. *Let $\mathcal{T} = (T, \phi)$ be a maximum parsimony X-tree for two characters χ_1 and χ_2. Then $\mathcal{T}$ can be a transformed by a series of edge contractions and rearrangements into a new maximum parsimony X-tree $\mathcal{T}' = (T', \phi')$ such that for every $v \in V(T')$, $\exists x \in X$ where $\phi'(x) = v$.*

Proof. Let $\bar{\chi_1}$ and $\bar{\chi_2}$ be two minimal extensions of χ_1 and χ_2 to $\mathcal{T}$ respectively. First create a new underlying tree T_0 by contracting every edge in $E(T) - (\mathrm{Ch}(\bar{\chi_1}, \mathcal{T}) \cup \mathrm{Ch}(\bar{\chi_2}, \mathcal{T}))$. Let $\mathcal{T}_0 = (T_0, \phi_0)$ be the corresponding

X–tree formed by T_0 and ϕ. Note that $l_{T_0}(\chi_1, \chi_2) = l_T(\chi_1, \chi_2)$ and that the minimal extensions for $\mathcal{T}$ can be mapped into minimal extensions for $\mathcal{T}_0$ as well.

Next, let $v \in V(T_0)$ be any vertex such that $\phi_0(x) \neq v, \forall x \in X$. Partition the set of adjacent vertices of v, N_v into three sets: $N_1 = \{u : \bar{\chi_1}(u) = \bar{\chi_1}(v)\}$, $N_2 = \{u : \bar{\chi_2}(u) = \bar{\chi_2}(v)\}$ and $N_3 = N_v - (N_1 \cup N_2)$. Note that N_1 and N_2 are disjoint. Remove v and its $|N_1| + |N_2| + |N_3|$ incident edges. Connect the vertices within each set N_1, N_2, N_3 to give three chains. Finally, let $a_1 \in N_1$, $a_2 \in N_2$ and $a_3 \in N_3$. Create two new edges (a_1, a_2) and (a_1, a_3) thereby creating a new underlying tree T' and corresponding X–tree $\mathcal{T}'$. Note that $\mathrm{ch}(\bar{\chi_1}, \mathcal{T}') \leq \mathrm{ch}(\bar{\chi_1}, \mathcal{T}_0) = \mathrm{ch}(\bar{\chi_1}, \mathcal{T})$ and $\mathrm{ch}(\bar{\chi_2}, \mathcal{T}') \leq \mathrm{ch}(\bar{\chi_2}, \mathcal{T}_0) = \mathrm{ch}(\bar{\chi_2}, \mathcal{T})$. Repeating this procedure for any such v completes the proof. □

Note that the series of rearrangements and contractions described in the previous lemma are not unique.

Theorem 2.4. *Let χ_1 and χ_2 be two characters defined on X. Let G be the complete graph on X with edges weights $w(x_1, x_2)$ defined as the Hamming distance between $(\chi_1(x_1), \chi_2(x_1))$ and $(\chi_1(x_2), \chi_2(x_2))$. Then any minimum weight spanning tree of G corresponds to a maximum parsimony X–tree for χ_1 and χ_2.*

Proof. Let T^* be the induced X–tree corresponding to a minimum weight spanning tree of G. Clearly $l_T(\chi_1, \chi_2) \leq l_{T^*}(\chi_1, \chi_2)$ where T is a maximum parsimony X–tree of χ_1 and χ_2. By applying the previous lemma, it is easy

to see T can be transformed into a tree that corresponds to a spanning tree of G showing that $l_{T^*}(\chi_1, \chi_2) \leq l_T(\chi_1, \chi_2)$ thereby completing the proof. $\square$

Author's Note

Subsequent to the submission of the present article a different algorithm for the two character case (similar to the spanning tree approach) was published independently by Althaus and Naujoks [1].

Acknowledgments

The authors would like thank the two anonymous reviewers for a number of helpful suggestions on the manuscript. TB acknowledges the support of a McGill Major scholarship.

References

[1] E. Althaus and R. Naujoks. Computing Steiner minimum trees in hamming metric. In *SODA '06: Proceedings of the seventeeth annual ACM-SIAM symposium on Discrete algorithms*, pages 172–181, 2006.

[2] T. C. Bruen, H. Philippe, and D. Bryant. A simple and robust statistical test to detect the presence of recombination. *Genetics*, 172:2665–2681, 2006.

[3] J. H. Camin and R. R. Sokal. A method for deducing branching sequences in phylogeny. *Evolution*, 19(3):311–326, 1965.

[4] G.F. Estabrook and L. Landrum. A simple test for the possible simultaneous evolutionary divergence of two amino acid positions. *Taxon*, 24:609–613, 1975.

[5] G.F. Estabrook and F. R. McMorris. When are two qualitative taxonomic characters compatible? *Journal of Mathematical Biology*, 4:195–200, 1977.

[6] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–76, 1981.

[7] L.R. Foulds and R. L. Graham. The Steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics*, 3:43–49, 1982.

[8] M. R. Garey and D. S. Johnson. *Computers and Intractibility.* W.H. Freeman, San Francisco, 1979.

[9] D. Gusfield. Efficient algorithms for inferring evolutionary trees. *Networks*, 21:19–28, 1991.

[10] B. R. Holland, K. T. Huber, D. Penny, and V. Moulton. The minmax squeeze: guaranteeing a minimal tree for population data. *Molecular Biology and Evolution*, 22:235–42, 2005.

[11] J. P. Huelsenbeck and F. Ronquist. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17:754–5, 2001.

[12] R. Karp. Reducibility among combinatorial problems. In R.E. Miller and J. W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–104. Plenum Press, 1972.

[13] D. Penny and M. Hendy. Estimating the reliability of evolutionary trees. *Molecular Biology and Evolution*, 3(5):403–17, 1986.

[14] C. Semple and M. A. Steel. *Phylogenetics.* Oxford University Press, 2003.

[15] M. A. Steel and D. Penny. Maximum parsimony and the phylogenetic information in multi-state characters. In V. Albert, editor, *Parsimony, phylogeny and genomics*, pages 163–178. Oxford University Press, 2005.

CHAPTER 3

Maximum parsimony is a consensus method

3.1 Background

This chapter presents another 'dual formulation' of maximum parsimony that links parsimony to a metric space on a set of trees. Direct links are provided between a standard formulation of maximum parsimony and a 'subtree-prune and regraft' (SPR) distance. Further links are shown between Dollo parsimony with the 'Robinson-Foulds' distance. The results have broad biological and mathematical impact.

3.2 Abstract

Matrix representation with parsimony is the most popular method to build supertrees. However the traditional interpretation of homoplasy makes little sense in this context. Here we show that maximum parsimony can be construed as a consensus method, where homoplasy has an elegant, alternative definition in terms of subtree-prune and regrafts. This motivates maximum parsimony as a consensus approach and explains the concept of homoplasy in the supertree setting. Two variants of Dollo parsimony have a similar relationship with the Robinson-Foulds and the Nearest-Neighbor

Interchange distances. We show that there are also connections between compatibility and tree space. The results give a new interpretation of maximum parsimony with special relevance as a supertree method.

3.3 Introduction

Matrix representation with parsimony (MRP) [2, 28, 3] is the most widely used method to build supertrees [4]. One of the main criticisms of the approach is that the traditional notion of homoplasy (e.g. convergent or recurrent mutation) in this context has no meaning [e.g. 30, 34, 37, 16, 11]. In particular, because of the lack of interpretability, it has been suggested that MRP should be treated as a 'black-box' approach [37]. We will show that MRP has an alternative motivation.

To begin with, let us consider the definition of a consensus tree (or supertree) both when the input consists of single trees and sets of trees. Suppose we have a set of input trees $T_1, ..., T_n$ and some measure of 'distance' or 'dissimilarity' between any two trees. Then a *median consensus tree* can be defined as a tree T that minimizes:

$$\sum_{i=1}^{n} d(T_i, T) = d(T_1, T) + d(T_2, T) + ... + d(T_n, T)$$

where $d(T, T_i)$ refers to the distance between tree T_i and tree T. The median consensus tree T then represents a centre tree of all the trees $T_1, ..., T_n$, under a dissimilarity measure d. In other words, T represents the closest tree to all the other trees.

Given that two trees are different, how should they be compared? A distance between trees can be taken as the Robinson-Foulds symmetric difference [29], or the smallest number of subtree prune and regrafts (SPRs) [e.g. 18] needed to transform one tree to another tree. For instance, if trees T_1 and T_2 could be transformed into each other using a minimum of 2 SPRs, $d_{SPR}(T_1, T_2)$ would equal two. The median consensus tree would then represent the tree that has the total smallest number of SPRs from all other trees. Defining a supertree as the tree that minimizes the SPR distance to each of the input trees is conceptually similar to the one proposed originally by Gordon [17].

But instead of individual trees, suppose the input may consist of a *set* or *group* of trees. For instance, individual studies may return many trees rather than a single tree. In this case, the goal is to combine each of these sets of trees rather than the individual trees themselves. Let $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_n$ each represent a set of trees. The 'dissimilarity' between the group of trees $\mathcal{S}_i$ and a single median consensus tree T can be defined as the *minimum* 'distance' between any tree in $\mathcal{S}_i$ and T, i.e.:

$$\sum_{i=1}^{n} d_m(\mathcal{S}_i, T) = d_m(\mathcal{S}_1, T) + ... + d_m(\mathcal{S}_n, T)$$

where d_m refers to the *minimum* 'distance' between any tree in $\mathcal{S}_i$ and T. Note that the choice of a *minimum* distance between $\mathcal{S}_i$ and T is somewhat arbitrary, an issue that we will return to later.

To see how sets of trees arise in the maximum parsimony context, consider a column in an multiple alignment representing a character X_j. Suppose there are six species $1, 2, 3, 4, 5$ and 6 and three different character states A, B and C. Suppose that species 1 and 2 share character state A, 3 and 4 state B and finally 5 and 6 share state C. Then there are a number of trees on which this character is homoplasy–free (Figure 3–1). Denote the set of trees on which character X_j is homoplasy-free by S_j.

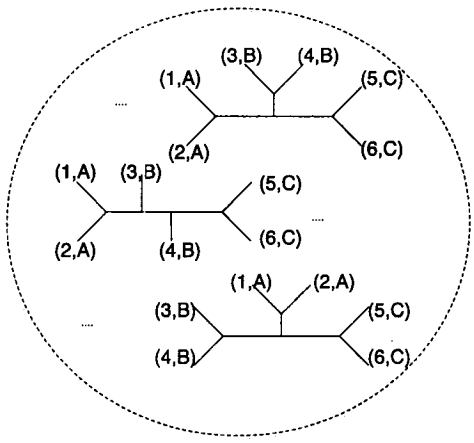


FIGURE 3–1: The character X_j on six species $1, 2, 3, 4, 5$ and 6 where both 1 and 2 are assigned A, 3 and 4 are assigned B, 5 and 6 are assigned C is homoplasy-free on many different trees. The leaves of each tree are labeled by each taxa and its corresponding character state. The set of all trees on which X_j is homoplasy-free is denote by S_j.

3.4 Main Result

We first state the main result and leave the proofs for the appendix. Note that the result holds for multi-state characters as well as binary characters.

Claim 3.1. *Given (multistate) characters $X_1, ..., X_n$, let $\mathcal{S}_i$ denote the set of trees on which character X_i is 'homoplasy-free'. Consider a median consensus tree (or supertree) T that minimizes the SPR distance to each of the sets of trees $\mathcal{S}_1, ..., \mathcal{S}_n$. The median consensus tree that results is the Fitch [15] maximum parsimony tree for characters $X_1, ..., X_n$. Furthermore, the homoplasy of each character on the optimal tree represents the minimimum number of subtree prune and regrafts (SPRs) needed to transform the optimal tree onto one of the trees that the character 'supports'.*

Informally stated, Claim 3.1 says that a maximum parsimony tree combines the trees 'implied' by each of the characters so that the number of SPRs is minimized (Figure 3–2). Note that missing data does not affect this result, and so this Claim extends to the case when some characters have missing states.

It turns out that this is not the only relationship between parsimony and distances between trees. In fact, Claim 3.1 is still valid if the word SPR is replaced by TBR, where the acronym TBR refers to the tree-bisection and reconnection distance (see the proof of results section). In the following section we state several other extensions of this result.

3.5 Extensions of main result

In the first Claim, we stated that minimizing the SPR (or TBR) distance to the set trees implied by each character was equivalent to (Fitch)

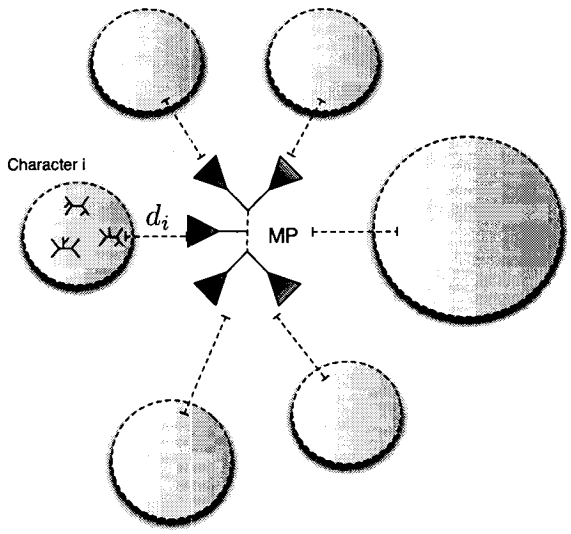


FIGURE 3–2: A maximum parsimony tree minimizes the homoplasy of all the characters. Each circle represents all the trees on which a specific character is homoplasy–free. In the case of Fitch parsimony, d_i is both the homoplasy of character i on T and is also the minimum number of SPRs between a maximum parsimony tree T and a tree on which character i is homoplasy-free. Two new variants of Dollo parsimony proposed in the text minimize a d_i that is similar to the RF and NNI distances respectively.

maximum parsimony. Here we consider what happens if we use other standard measures for comparing trees.

3.5.1 Discrete distances and maximum compatibility

Perhaps the simplest measure of distinctiveness between two trees is the *discrete distance* defined by $d_0(T_1, T_2) = 0$ if T_1 and T_2 refer to an identical tree, and 1 otherwise. If the input consists of the sets of trees on which each character is 'homoplasy-free', then the median consensus tree that minimizes the discrete distance is the maximum compatibility tree [24] (that is the

The Dollo property states that the derived state 1 must be uniquely derived [13, 25]. In particular, no convergent mutations or reversals with respect to the derived state 1 are allowed. The maximum Dollo parsimony tree T minimizes the number of changes for all binary characters on T, where each character is subject to the Dollo property. Since it has been shown that the Dollo parsimony score is independent of the placement of the root [36, 22], it is more convenient to consider unrooted trees rather than rooted trees. The Dollo parsimony score for an unrooted tree T can be thought of as the score for any arbitrarily rooted version of T.

Consider a binary character X_j with character states 0 and 1. Define a *complementary character* $\bar{X}_j$ with the property that $\bar{X}_j$ inverts the derived and ancestral states. It turns out that finding the maximum Dollo parsimony tree for all characters X_j and $\bar{X}_j$ is similar to finding the tree that minimizes the RF distance to 'groups' of trees, where each 'group' of trees represent the trees on which a given character is homoplasy free. This is stated precisely in the following Claim:

Claim 3.2. *Given (binary) characters $X_1, ..., X_n$ and their complements $\bar{X}_1, ..., \bar{X}_n$, let $\mathcal{S}_i$ denote the set of trees on which character X_i (and hence $\bar{X}_1$) is 'homoplasy-free'. Consider a median consensus tree (or supertree) T that minimizes the sum of the RF and discrete distance to each of the sets of trees $\mathcal{S}_1, ..., \mathcal{S}_n$. The median consensus tree that results is the Dollo maximum parsimony tree for characters $X_1, \bar{X}_1, ..., X_n, \bar{X}_n$.*

The relationship of RF distance and Dollo parsimony can be extended to give relationship between the NNI distance and Dollo parsimony. Essentially, with a modest adjustment in the total cost, by counting the nodes in the tree that strongly conflict with the characters an analogue to Claim 3.2 for the NNI distances can be developed. The result is further detailed in the proof of results section.

3.6 Two character incompatibility and recombination

As a special case, we consider two multistate characters. Two binary characters with states 0 and 1 are *incompatible* if and only if all four combinations of $00, 01, 10$, and 11 are present within the taxa [24]. In a standard setting, character incompatibility implies that at least one of the characters has undergone convergent or recurrent mutation (homoplasy). In the absence of recurrent or convergent mutation, incompatibility of two sites implies at least one recombination event has occurred between both sites [35, 21].

We have recently shown how to calculate a *refined incompatibility* score efficiently [7]. Instead of two characters being either compatible or incompatible, refined incompatibility considers the degree of incompatibility. For instance a refined incompatibility score of 2 for two characters indicates at least two recurrent or convergent mutations are needed on any tree with both characters. Although this notion of refined incompatibility has been considered before in the context of character selection and weighting [27], it has not been considered in the context of recombination. Essentially

Claim 3.3 shows that in the absence of recurrent or convergent mutation (i.e. homoplasy), the refined incompatibility score can be interpreted as the minimum number of recombinations that have occurred between two sites. Indeed, Claim 3.3 suggests a natural way to interpret incompatibility between two characters, which we have used to develop a powerful test for recombination [8].

Claim 3.3. *Let X_1 and X_2 be two multistate unordered characters with r_1 and r_2 states respectively that have maximum (Fitch) parsimony score $l(X_1, X_2)$ (over all trees) and refined incompatibility $i(X_1, X_2) = l(X_1, X_2) - r_1 - r_2 + 2$. Then $i(X_1, X_2)$ equals the minimum SPR distance $d_{SPR}(T_1, T_2)$ where T_1 and T_2 are any two trees on which X_1 and X_2 are homoplasy-free.*

Claim 3.3 shows that encoding two trees by five or four multi-state characters each [32, 20] a lower bound on the SPR distance between both trees can also be determined. This bound consists of calculating the refined incompatibility score between all pairs of characters where one of the characters is used to encode the first tree and the other character is used to encode the second tree. This is also a lower bound for the TBR distance, which is known to be computationally hard [1].

3.7 Discussion

In a consensus approach, input trees that have the same leaf set are combined into a consensus tree [9]. The criteria chosen to combine the trees is meant to capture the similarity between the input trees. Imagine combining sets of trees instead of single trees, so that the consensus tree minimizes

the SPR distance to each of the sets of trees. Then this is precisely maximum parsimony! In the parsimony case, the set of trees represent the trees on which each character is homoplasy-free. This alternative motivation of parsimony gives an new interpretation of homoplasy. The homoplasy of a single character on a tree T represents the minimum SPR distance between any tree on which the character is homoplasy-free and T (Figure 3–2). This is quite different than the standard interpretation of homoplasy, where the homoplasy of character on a tree T represents the minimum number of convergent or recurrent mutations of the character on T [36].

In fact, the 'consensus' interpretation of maximum parsimony gives possible reasons for the success of MRP [6], but also exposes some of its shortcomings [see also 5]. Clearly, MRP partially succeeds as a supertree approach because parsimony is in some sense a consensus method! On the other hand, MRP only indirectly combines input trees through the mechanisms of character encoding. Moreover, it chooses a tree that is 'minimally' distant from all trees supported by that character. One question that arises is whether the choice of minimum distance is necessarily optimal. For instance, it may be better to take a sum or median distance. More directly though, another question that arises is the effect of character encoding, for instance binary characters or multi-state characters, or even character weighting. Instead of encoding trees as characters, ideally, the trees themselves would be directly combined under some criteria [e.g. 23, 34]. Indeed, combining input trees so that the output tree minimizes

the SPR distance (representing lateral gene transfers) has been suggested in a gene tree context [26]. But combining input trees directly, so that the output tree minimizes the SPR distance to the input trees for instance, is likely to be computationally difficult since even computing the SPR distance between two trees is likely computationally hard [19, 1]. However, our work suggest a simple if crude way to obtain lower bounds for this problem.

The conception of parsimony as a consensus method does not remove its shortcomings as a general approach to infer phylogenies [e.g. 14]. Nonetheless, it provides a new paradigm to think about maximum parsimony and the trees that are inferred using the approach. Moreover, it provides valuable insight that explains the concept of homoplasy in the context of matrix representation with parsimony.

Acknowledgements

TB would like to thank M. Steel for some valuable comments on the manuscripts.

References

[1] B. Allen and M. A. Steel. Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, 5:1–13, 2001.

[2] B. R. Baum. Combining trees as a way of combining datasets for phylogenetic inference, and the desirability of combining gene trees. *Taxon*, 41:3–10, 1992.

[3] B.R. Baum and M. A. Ragan. *Phylogenetic supertrees: Combining information to reveal the tree of life*, chapter The MRP Method, pages 17–34. Kluwer Academic Press, 2004.

[4] O. R. Bininda-Emonds. The evolution of supertrees. *Trends in Ecology and Evolution*, 19(6):315–322, 2004.

[5] O. R.. Bininda-Emonds, editor. *Phylogenetic Supertrees.* Kluwer Academic Publishers, 2004.

[6] O. R. Bininda-Emonds and M. J. Sanderson. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. *Systematic Biology*, 50(1063-5157):565–79, 2001.

[7] T. C. Bruen and D. Bryant. A subdivision approach to maximum parsimony. *Annals of Combinatorics*, In Press, 2006.

[8] T. C. Bruen, H. Philippe, and D. Bryant. A simple and robust statistical test for detecting the presence of recombination. *Genetics*,

172:2665–2681, 2006.

[9] D. Bryant. A classification of consensus methods for phylogenetics. In *Bioconsensus*, volume 61 of *DIMACS*, pages 163–184. American Math Society, Providence, RI, 2003.

[10] D. Bryant. The splits in the neighborhood of a tree. *Annals of Combinatorics*, 8(1):1–11, 2004.

[11] H. N. Bryant. *Phylogenetic supertrees: Combining information to reveal the tree of life*, chapter The cladistics of matrix representation with parsimony analysis, pages 353–369. Kluwer Academic Publishers, 2004.

[12] P. Buneman. The recovery of trees from measure of dissimilarity. In D.G. Kendall F.R. Hodson and P. Tautu, editors, *Mathematics in the Archaelogical and Historical Sciences*. Edinburgh University Press, Edinburgh, pp. 387-395, 1971.

[13] J. S. Farris. Phylogenetic analysis under Dollo's law. *Systematic Zoology*, 26:77–88, 1977.

[14] J. Felsenstein. Cases in which parsimony and compatibility methods will be positively misleading. *Systematic Zoology*, 27:401–410, 1978.

[15] W. M. Fitch. Towards defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, 20(4):406–416, 1971.

[16] J. Gatesy and M. S. Springer. *Phylogenetic supertrees: Combining information to reveal the tree of life*, chapter A critique of matrix representation with parsimony supertrees, pages 369–388. Kluwer

Academic Press, 2004.

[17] A. D. Gordon. Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. *Journal of Classification*, 3:335–348, 1986.

[18] J. Hein. Reconstructing evolution of sequences subject to recombination using parsimony. *Mathematical Biosciences*, 98(2):185–200, 1990.

[19] J. Hein, T. Jiang, L. Wang, and K. Zhang. On the complexity of comparing evolutionary trees. *Discrete Applied Mathematics*, 71(1-3):153–169, 1996.

[20] K. T. Huber, V. Moulton, and M. A. Steel. Four characters suffice to convexly define a phylogenetic tree. *SIAM Journal on Discrete Mathematics*, 18(4):835–843, 2005.

[21] R. R. Hudson and N. L. Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111(1):147–64, 1985.

[22] D. H. Huson and M. A. Steel. Phylogenetic trees based on gene content. *Bioinformatics*, 20:2044–9, 2004.

[23] F. Lapointe and G. Cucumel. The average consensus procedure: combination of weighted taxa containing identical or overlapping sets of taxa. *Systematic Biology*, 46(2):306–312, 1997.

[24] W. J. Le Quesne. A method of selection of characters in numerical taxonomy. *Systematic Zoology*, 18(2):201–205, 1969.

[25] W. J. Le Quesne. The uniquely evolved character concept and its cladistic application. *Systematic Zoology*, 23(4):513–517, 1974.

[26] W. P. Maddison. Gene trees in species trees. *Systematic Biology*, 46(3):523–536, 1997.

[27] D. Penny and M. Hendy. Estimating the reliability of evolutionary trees. *Molecular Biology and Evolution*, 3(5):403–17, 1986.

[28] M. A. Ragan. Phylogenetic inference based on matrix representations of trees. *Molecular Phylogenetics and Evolution*, 1:53–58, 1992.

[29] D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.

[30] A. Rodrigo. A comment on Baum's method for combining phylogenetic trees. *Taxon*, 42:631–636, 1993.

[31] H. A. Ross and A. G. Rodrigo. *Phylogenetic supertrees: Combining information to reveal the tree of life*, chapter An assessment of matrix representation with compatibility in supertree construction, pages 35–63. Kluwer Academic Press, 2004.

[32] C. Semple and M. A. Steel. Tree reconstruction from multi-state characters. *Advances in Applied Mathematics*, 28(2):169–84, 2002.

[33] C. Semple and M. A. Steel. *Phylogenetics*. Oxford University Press, 2003.

[34] J. B. Slowinski and Roderic D. M. Page. How should species phylogenies be inferred from sequence data. *Systematic Biology*, 48(4):814–825, 1999.

[35] P. H. A. Sneath, M. J. Sackin, and R. P. Ambler. Detecting evolutionary incompatibilities from protein sequences. *Systematic Zoology*, 24(3):311–332, 1975.

[36] D. L. Swofford, G. J. Olsen, P. J. Wadell, and D. M. Hillis. *Molecular sytematics*, chapter Phylogenetic Inference, pages 407–514. Sinauer Associates, Inc., 1996.

[37] M. Wilkinson, J. L. Thorley, D. T. J. Littlewood, and R. A. Bray. *Interrelationships of the Platyhelminthes*, chapter Towards a phylogenetic supertree of Platyhelminthes?, pages 292–301. Taylor and Francis, 2001.

3.9 Proof of Results

We refer the interested reader to the book of Semple and Steel for a description of the notation [33]. Additionally, for a character χ, let S_χ denote the set of binary phylogenetic X-trees on which χ is convex.

3.9.1 Main result

We first restate a Lemma of [10].

Lemma 3.1. [10] *Let T be a phylogenetic X–tree and χ a multistate character. Let T' be phylogenetic X tree that differs from T by a single TBR. Then $l(\chi, T') \leq l(\chi, T) + 1$.*

The version of Lemma 3.1 that is proved in [10] (Lemma 5.1) is stated for binary characters but applies to multistate characters as well. The following Lemma also expands on results of [10].

Lemma 3.2. *Let T be a binary phylogenetic X–tree and χ a multistate character. Then we have the following relationship $h(\chi, T) = \min_{T' \in S_\chi} d_{SPR}(T, T')$, where $h(\chi, T)$ denotes the homoplasy of χ on T.*

Proof. Let T' be any binary phylogenetic X–tree for which $d_{SPR}(T, T') = m$ is minimized and χ is convex on T'. Then there exists a sequence of trees $T' = T_0, ..., T_m = T$ such that every adjacent pair of trees in the sequence differ by exactly one SPR. Since χ is convex on T' and in particular every SPR is a TBR, then by Lemma 3.1 the existence of this sequence implies that $h(\chi, T) \leq d_{SPR}(T, T')$.

For the other direction, we need to show that $\min_{T' \in S_\chi} d_{SPR}(T, T') \leq h(\chi, T)$. To do this we will construct a sequence of trees $T_0 = T, ..., T_k$ such that every pair of adjacent trees in the sequence differ by exactly one SPR and χ is convex on T_k where $k = h(\chi, T)$. Firstly, if $h(\chi, T) = 0$, then χ is convex on T so the proof is finished. Otherwise, let $\bar{\chi}$ be a minimum length extension of χ to T. Then since χ is not convex on T there exist three vertices u, v and w, where $\{u, v\} \in E(T)$, v lies on the path from u to w and $\bar{\chi}(u) = \bar{\chi}(w) \neq \bar{\chi}(v)$. Then perform a SPR by removing edge $\{u, v\}$, supressing the v vertex and creating a new edge $\{u, t\}$ where t is a new vertex on an edge adjacent to w. Furthermore, set $\bar{\chi}(t) = \bar{\chi}(w)$. Then the number of edges on which a change has occurred has decreased by 1 thereby decreasing the homoplasy score by 1. This procedure can be repeated until the homoplasy equals 0, constructing the desired sequence of trees thereby demonstrating $\min_{T' \in S_\chi} d_{SPR}(T, T') \leq k = h(\chi, T)$ and completing the proof. □

We now state the main Theorem.

Theorem 3.1. *Consider a sequence $\mathcal{C} = (\chi_1, ..., \chi_n)$ of unordered, multi-state characters where $\chi_i : X \to C$ for all i. Then T is a binary, phylogenetic (classical) maximum parsimony X-tree for $\mathcal{C}$ if and only if $\sum_{j=1}^{n} d_{SPR}(T, T_j)$, $T_j \in S_{\chi_j}$ is minimized. Moreover, the homoplasy of character χ_j on T refers to the minimum number of SPRs needed to transform T into a tree T_j on which χ_j is convex.*

The proof of Theorem 3.1 follows directly by Lemma 3.2.

3.9.2 *Dollo parsimony and related results*

First, some additional notation. For any binary character $\chi : X \to \{0,1\}$ define a *complementary character*, $\hat{\chi} : X \to \{0,1\}$ by $\hat{\chi}(x) = 0$ if and only if $\chi(x) = 1$ and $\hat{\chi}(x) = 1$ if and only if $\chi(x) = 0$.

We also require some additional notation from Bryant [10]. An edge of a phylogenetic X-tree T that induces a split $C|D$ is said to *conflict* with another split $A|B$ if none of $A \cap C, A \cap D, B \cap C$ and $B \cap D$ are empty. A vertex v of $V(T)$ is said to conflict with $A|B$ if all three incident edges conflict with $A|B$.

Furthermore, for a metric d defined on the set of binary phylogenetic X-trees, $UB(X)$, the *r-neighbourhood* of T with respect to d equals the set of trees

$$N_d(T,r) = \{T' \in UB(X) : d(T,T') \leq r\}$$

The *split neighborhood* of T is the set of splits appearing in at least one of the trees in the r neighborhood of T:

$$S_d(T,r) = \{A|B : \text{ there exists } T' \in N_d(T,r) \text{ such that } A|B \in \Sigma(T)\}$$

We recall a Theorem from [10].

Theorem 3.2. *[10] Let T be a fully resolved phylogenetic X-tree. A split $A|B$ is in $S_{RF}(T,r)$ if and only if $A|B$ conflicts with at most r edges of T.*

For our purposes, we need a corollary of Theorem 3.2 as a Lemma.

Lemma 3.3. *Let χ be a binary character that corresponds to a split $A|B$ (e.g. $\chi(A) = 0$ and $\chi(B) = 1$), and T a binary phylogenetic X–tree. Let r be the number of edges of T that conflict with $A|B$. Then $\min_{T' \in S_\chi} d_{RF}(T, T') = r$.*

Proof. Suppose $A|B$ conflicts with r edges of T. By Theorem 3.2 there exists a binary phylogenetic X-tree T' for which $d_{RF}(T, T') \leq r$, $A|B \in \Sigma(T')$ and hence χ is convex on T'. On the other hand, suppose that $d_{RF}(T, T')$ is minimized for any tree $T' \in S_\chi$. Then χ is convex on T' and so $A|B \in \Sigma(T')$. Then by Theorem 3.2, $A|B$ conflicts with at most s edges of T, i.e. $r \leq d_{RF}(T, T')$. □

Next we characterize the set of conflicting of edges. We also adopt some additional notation adopted from Huson and Steel [22]. Consider an unrooted phylogenetic X-tree T, and a binary character $\chi : X \to \{0, 1\}$. Let $p(x, y) := p_T(x, y)$ denote the set of vertices on the path in T connecting x and y. Let

$$V(\chi, T) = \{v \in V(T) : \exists x, y \in X : \chi(x) = \chi(y) = 1, v \in p(x, y)\},$$
$$E(\chi, T) = \{\{u, v\} \in E(T) : u \in V(\chi, T) \text{ and } v \in V(\chi, T)\},$$

and

$$\Delta(\chi, T) = |\{\{u, v\} \in E(T) : |\{u, v\} \cap V(\chi, T)| = 1\}|.$$

Consider a rooted phylogenetic X-tree T (with edges directed away from the root), and a binary character $\chi : X \to \{0, 1\}$. The *DP score* of χ

on T, $l(\chi, T)$ is the minimal length of a *DP extension* $\bar{\chi}$ for which: i) there is at most one edge (u, v) in T with $\bar{\chi}(u) = 0$ and $\bar{\chi}(v) = 1$, and ii) the number of edges (u, v) such that $\bar{\chi}(u) \neq \bar{\chi}(v)$ is minimized. Note that if $e \in \Delta(\chi, T)$ then at least on of the components of $T \backslash \Delta(\chi, T)$ is monochromatic with respect to the coloring induced by χ (the component that contains no vertices labeled with 1).

To begin with we recall a Lemma of Huson and Steel [22]. For a rooted phylogenetic X–tree T, let $T^{-\rho}$ denote the unrooted phylogenetic X-tree obtained by suppressing the root vertex ρ. Moreover let $l_{DP}(\chi, T)$ be the DP-score of χ on T. Their result allows us to consider unrooted trees rather than rooted trees.

Lemma 3.4. [22] *For a rooted phylogenetic X-tree T and a character $\chi : X \to \{0, 1\}$ we have*

$$l_{DP}(\chi, T) = \Delta(\chi, T^{-\rho})$$

Thus, $l_{DP}(\chi, T)$ is independent of the placement of a root.

Then for an *unrooted* phylogenetic X–tree T, we can define the DP score for T as $l_{DP}(\chi, T^{+\rho})$, where $T^{+\rho}$ is an arbitrary rooting of T. By Lemma 3.4, this notion is well-defined. The following Lemma begins to connect the Dollo parsimony score of a binary character on a tree with the conflicting edges of a split.

Lemma 3.5. *Let $\chi : X \to \{0,1\}$ and $\hat{\chi} : X \to \{0,1\}$ be two complementary characters that correspond to a split $A|B$ and T a phylogenetic X-tree. Then $e \in E(T,\chi) \cap E(T,\hat{\chi})$ if and only if e conflicts with $A|B$.*

Proof. Suppose without loss of generality that $\chi(A) = 1 = \hat{\chi}(B)$ and $\chi(B) = 0 = \hat{\chi}(A)$. Let $e \in E(T)$ be an edge of T and denote the split that e induces by $C|D$. Suppose that e conflicts with $A|B$, (note that e is forcibly an interior edge). Then by definition since e conflicts with $A|B$, there $\exists a \in C \cap A$ and $\exists b \in D \cap A$. Then by supposition, $\chi(a) = \chi(b) = 1$, implying that $e \in E(T,\chi)$. Similarly, we can show that $e \in E(T,\hat{\chi})$.

On the other hand, suppose that $e \in E(T,\chi) \cap E(T,\hat{\chi})$. Then since $e \in E(T,\chi)$, there $\exists a \in C$ and $\exists b \in D$ such that $\chi(a) = \chi(b) = 1$. But $\chi(a) = 1$ if and only if $a \in A$, so $a \in A \cap C$ and $b \in A \cap D$. Similarly, we can show that $B \cap C$ and $B \cap D$ are both non-empty showing that e conflicts with $A|B$. □

We also need a small counting Lemma (short proof courtesy of M. Steel personal communication).

Lemma 3.6. *Let $T = (V,E)$ be a unrooted tree such that $deg(v) \leq 3$ for all $v \in V$ and $|E| \geq 1$. Let n_1 and n_2 be the number of vertices of degree 1 and 2 respectively. Then $2n_1 + n_2 = |E| + 3$.*

Proof. In any graph, the sum of the degree of the vertices is equal to twice the number of edges. Hence, $n_1 + 2n_2 + 3n_3 = 2|E|$. Next since T is a tree

$|E| = |V| - 1 = n_1 + n_2 + n_3 - 1$. Combining these two equations to eliminate n_3 gives us the desired relationship namely $2n_1 + n_2 = |E| + 3$. □

Lemma 3.7. *[10] If T is a full resolved X-tree and $A|B$ is a split of X then the edges of T conflicting with $A|B$ form a connected subgraph of T.*

We also restate a fundamental result in mathematical phylogenetics, known as the 'Splits Equivalence Theorem' [12, 33].

Theorem 3.3. *[12, 33] Let Σ be a collection of X-splits. Then, there is an X-tree T such that $\Sigma = \Sigma(T)$ if and only if the splits in Σ are pairwise compatible. Moreover, if such an X-tree exists, then, up to isomorphism, T is unique.*

The 'Splits Equivalence Theorem' leads to the following result, which was stated without proof in [10].

Lemma 3.8. *Let T be a binary phylogenetic X-tree, with splits $\Sigma(T)$. Then $\Sigma(T)$ is maximal set of pairwise compatible splits.*

Proof. By Theorem 3.3, the splits in $\Sigma(T)$ are pairwise compatible. Suppose $\Sigma(T)$ are not maximal, and consider a maximal set of pairwise compatible splits Σ' such that $\Sigma(T) \subset \Sigma'$. By Theorem 3.3, there is a phylogenetic X-tree T' such that $\Sigma(T') = \Sigma'$. Note that we may assume every internal vertex v of T' has $\deg(v) \geq 3$ since any degree two vertex can be contracted resulting in a tree with an identical set of splits. If for some vertex v', $\deg(v') \geq 4$ then v' can be replaced by v_1 and v_2 along with an edge $\{v_1, v_2\}$ so that $\deg(v_1) \geq 3$ and $\deg(v_2) \geq 3$. But the presence of the additional split induced by $\{v_1, v_2\}$ contradicts the maximality of Σ'. So T' is a binary

phylogenetic X–tree and hence $|\Sigma(T)| = |\Sigma'|$ a contradiction. Hence $\Sigma(T)$ is maximal. □

We also need another technical Lemma.

Lemma 3.9. *Let $\chi : X \to \{0,1\}$ and $\hat{\chi} : X \to \{0,1\}$ be two complementary characters that correspond to a split $A|B$ and T a binary phylogenetic X-tree. Then if χ (and hence $\hat{\chi}$) is not convex on T, the set of edges S that conflict with $A|B$ from a connected subgraph of T, with $e \in S$ if and only if $e \in E(\hat{\chi}, T) \cap E(\chi, T)$. Moreover if χ (and hence $\hat{\chi}$) is not convex on T, then $\Delta(\chi, T) \cap \Delta(\hat{\chi}, T) = \emptyset$ so that $|\Delta(\chi, T) \cup \Delta(\hat{\chi}, T)| = |\Delta(\chi, T)| + |\Delta(\hat{\chi}, T)|$*

Proof. Suppose χ is not convex on T. Then there is no edge whose removal induces the split $A|B$, and so $A|B \notin \Sigma(T)$. Since T is a binary phylogenetic X-tree, $\Sigma(T)$ is maximal by Lemma 3.8. So $A|B$ conflicts with a least one edge of T and by Lemma 3.7 the set of edges that conflict with $A|B$ form a (non-empty) connected subgraph. By Lemma 3.5, there is bijection between the set of edges that conflict with $A|B$ and $E(T, \chi) \cap E(T, \hat{\chi})$. Note that $\Delta(\chi, T) \cap \Delta(\hat{\chi}, T) = \emptyset$, since otherwise if $e \in \Delta(\chi, T) \cap \Delta(\hat{\chi}, T)$ removal of e gives two monochromatic components with respect to the coloring induced by χ and $\hat{\chi}$, implying that χ is convex on T. □

Lemma 3.10. *Let $\chi : X \to \{0,1\}$ and $\hat{\chi} : X \to \{0,1\}$ be two complementary characters that correspond to a split $A|B$ and T a binary phylogenetic X-tree. Then if χ (and hence $\hat{\chi}$) is not convex on T, an edge $e \in \Delta(\chi, T) \cup \Delta(\hat{\chi}, T)$ if and only if e is incident to an edge $f \in E(\hat{\chi}, T) \cap E(\hat{\chi}, T)$.*

Proof. If χ is not convex then by Lemma 3.9 then $E(T,\chi) \cap E(T,\hat{\chi})$ forms a non-empty connected set of edges that conflict with χ.

Suppose e is an edge incident to $E(T,\chi) \cap E(T,\hat{\chi})$. Then this implies that $e \in \Delta(\chi,T) \cup \Delta(\hat{\chi},T)$ because either $e \notin E(T,\chi)$ or $e \notin E(T,\hat{\chi})$. On the other hand suppose that $e \in \Delta(\chi,T) \cup \Delta(\hat{\chi},T)$. Then by Lemma 3.9 since χ is not convex on T these two sets are disjoint so suppose without loss of generality that $e \in \Delta(\chi,T)$. Let $e = \{u,v\}$ where $v \in V(\chi,T)$ (and hence $u \in V(\hat{\chi},T)$). Denote the other two edges incident to v by e_1 and e_2. Note that both $e_1 \in E(T,\chi)$ and $e_2 \in E(T,\chi)$ since otherwise $v \notin V(\chi,T)$. Note also that either $e_1 \in E(T,\hat{\chi})$ or $e_2 \in E(T,\hat{\chi})$ since otherwise $v \notin V(\hat{\chi},T)$ and so $e \in \Delta(\hat{\chi},T)$ a contradiction. Hence e is incident to an edge in $E(T,\chi) \cap E(T,\hat{\chi})$. □

The main Lemma. Let $h_{DP}(\chi,T) = l_{DP}(\chi,T) - 1$ denote the *Dollo homoplasy* of a binary character $\chi : X \to \{0,1\}$ on a phylogenetic X-tree T.

Lemma 3.11. *Let $\chi : X \to \{0,1\}$ and $\hat{\chi} : X \to \{0,1\}$ be two complementary characters and T a phylogenetic X-tree. Then if χ (and hence $\hat{\chi}$) is not convex on T, $h_{DP}(\chi,T) + h_{DP}(\hat{\chi},T) = \min_{T' \in S_\chi} (d_{RF}(T,T') + d_0(T,T'))$, where d_{RF} refers to the RF distance and d_0 refers to the discrete distance.*

Proof. Let $A|B$ be the split of X for which $\chi(A) = 1 = \hat{\chi}(B)$ and $\chi(B) = 0 = \hat{\chi}(A)$. If χ (and hence $\bar{\chi}$) is convex on T then clearly the statement is proved. Suppose that χ is not convex on T. By Lemma 3.9 the set of edges S that conflict with $A|B$ forms a connected subgraph with $e \in S$ if and only if $e \in E(\hat{\chi},T) \cap E(\hat{\chi},T)$. Then Lemma 3.3 gives the

following relationship: $\min_{T' \in S_\chi} d_{RF}(T, T') = |E(\chi, T) \cap E(\chi, T)| = |S|$. Note that the cardinality of the edge set incident to the connected subgraph $S = E(\hat{\chi}, T) \cap E(\hat{\chi}, T)$ is precisely $2n_1 + n_2$ where n_1 and n_2 consist of the degree one and degree two vertices in S. By Lemma 3.6, we have that $2n_1 + n_2 = |S| + 3$. Next note that by Lemma 3.10, the set of edges that are incident to S is precisely $\Delta(\chi, T) \cup \Delta(\hat{\chi}, T)$. Since by Lemma 3.9 $\Delta(\chi, T)$ and $\Delta(\hat{\chi}, T)$ are disjoint, we have that $|\Delta(\chi, T)| + |\Delta(\hat{\chi}, T)| = |S| + 3$ and so $|\Delta(\chi, T)| + |\Delta(\hat{\chi}, T)| = \min_{T' \in S_\chi} d_{RF}(T, T') + 3$. Finally by Lemma 3.4 $|\Delta(\chi, T)| + |\Delta(\hat{\chi}, T)| = h_{DP}(\chi, T) + h_{DP}(\hat{\chi}, T) + 2$ and so noting that $d_0(T', T) = 1$ if and only if χ is not convex on T we have that $h_{DP}(\chi, T) + h_{DP}(\hat{\chi}, T) = \min_{T' \in S_\chi} (d_{RF}(T, T') + d_0(T, T'))$. □

The following Theorem then follows directly from Lemma 3.11:

Theorem 3.4. *Consider a sequence $(\chi_1, ..., \chi_n)$ unordered, binary characters and their complements $(\bar{\chi_1}, ..., \bar{\chi_n})$ where $\chi_i : X \to C$ for all i. Then T is a maximum (Dollo) parsimony tree for $\mathcal{C}$ if and only if $\sum_{j=1}^{n} d_{RF}(T, T_j) + d_0(T, T_j)$, $T_j \in S_\chi$ is minimized. Moreover, the sum of the homoplasy of character χ_j and $\bar{\chi_j}$ on T refers to the minimum sum of the RF and discrete distance needed to transform T into a tree T_j on which χ_j (and hence $\bar{\chi_j}$) is convex.*

Theorem 3.4 can be directly extended using the NNI distance rather than RF distance by considering conflicting vertices. Essentially, by extending the Dollo parsimony score to count conflicting vertices as well, an

analogue of Theorem 3.4 for NNI distance follows. The result depends on Theorem 4.1 of Bryant's [10].

3.9.3 Two character incompatibility

Let $i(\chi_1, \chi_2) = l(\chi_1, \chi_2) - (|\chi_1| + |\chi_2|) + 2$ denote the *pairwise homoplasy* or *incompatibility* of χ_1 and χ_2 where $l(\chi_1, \chi_2)$ refers to the maximum parsimony score taken over all trees.

Theorem 3.5. *Let χ_1 and χ_2 be two multistate unordered characters with r_1 and r_2 states respectively that have maximum parsimony score $l(\chi_1, \chi_2)$ and incompatibiliy $i(\chi_1, \chi_2) = l(\chi_1, \chi_2) - r_1 - r_2$. Then $i(\chi_1, \chi_2) = \min_{(T_1, T_2) \in S_{\chi_1} \times S_{\chi_2}} d_{SPR}(T_1, T_2)$*

Proof. Let T be a maximum parsimony phylogenetic X-tree for χ_1 and χ_2, such that $h(\chi, T) = k$. By Lemma 3.2 there exist two trees T_1 and T_2 on which χ_1 and χ_2 are convex respectively such that $d_{SPR}(T_1, T) = k_1$ and $d_{SPR}(T_2, T) = k_2$, with $k_1 + k_2 = k$. Then by concatenating the sequence of trees from T_1 to T with T to T_2, we obtain $d_{SPR}(T_1, T_2) \leq k_1 + k_2 = k$. To complete the proof, let T_1' and T_2' be any two trees on which χ_1 and χ_2 are convex and $d_{SPR}(T_1', T_2')$ is minimized. Then $h_(\chi_1, T_1') = 0$ and so $h(\chi_2, T_1') \geq k_1 + k_2 = k$ (since T is a maximum parsimony tree) and so by Lemma 3.2, $k = k_1 + k_2 \leq d_{SPR}(T_1', T_2')$, completing the proof.

□

CHAPTER 4

A simple and robust statistical test for detecting recombination

4.1 Background

Determining whether or note recombination has occurred is an important biological question. A statistical test for detecting recombination is developed based on the observation that in the case of recombination, nearby sites will tend to share greater evolutionary similarity than distant sites. Similarity between pairs of sites can be calculated using a modified version of the two character parsimony score developed in Chapters 2 and 3.

4.2 Abstract

Recombination is a powerful evolutionary force that merges historically distinct genotypes. But the extent of recombination within many organisms is unknown, and even determining its presence within a set of homologous sequences is a difficult question. Here we develop a new statistic, Φ_w, that can be used to test for recombination. We show through simulation that our test can effectively discriminate between the presence and absence of recombination, even in diverse situations such as exponential growth

(starlike topologies) and patterns of substitution rate correlation. A number of other tests, Max χ^2, NSS, a coalescent based likelihood permutation test (from LDHat), and correlation of linkage disequilibrium (both r^2 and $|D'|$) with distance, all tend to underestimate the presence of recombination under strong population growth. Moreover, both Max χ^2 and NSS falsely infer the presence of recombination under a simple model of mutation rate correlation. Results on empirical data show that our test can be used to detect recombination between closely as well as distantly related samples, regardless of the suspected rate of recombination. The results suggest that Φ_w is one of the best approaches to distinguish recurrent mutation from recombination in a wide variety of circumstances.

4.3 Introduction

Recombination is a fundamental biological process that can, for example, increase viral or bacterial pathogenicity by diffusing genetic material throughout populations [2]. The biological mechanisms of recombination differ across organisms, but in broad terms recombination results in the creation of mosaic sequences where the evolutionary history at each site may be different. Violating this tree–like assumption of evolution can lead to serious consequences when performing phylogenetic analyses for a set of sequences. Indeed, as the evolution of the sequences cannot be described by a single tree, this can lead to overestimation or underestimation of branch lengths among other problems [64, 65, 56, 59]. Thus, an important question for a

given set of aligned sequences is to determine whether or not recombination is likely to have occurred.

The ability of a large number of general methods to detect recombination has recently been evaluated empirically and through simulation [9, 4, 58, 79, 57]. These studies have established that methods such as Geneconv [62], Max χ^2 [43], RDP [40], Phypro [77], RecPars [19, 20] and NSS [30] efficiently detect recombination in a wide range of circumstances [4, 58, 79, 57]. These tests infer the presence of recombination either directly through sequence comparisons, or indirectly through phylogenetic means. As no underlying assumptions are made concerning the origin of the sequences, these tests can be applied to detect recombination within any set of aligned homologous sequences. Indeed, these techniques can be used to detect recombination within either closely or distantly related genotypes [57]. Moreover, these methods for detecting recombination can be termed *general* since no specific assumptions concerning sample history (beyond sequence homology) are made.

In contrast to general methods for inferring recombination, there are also population specific methods for detecting recombination, where the samples consist of genotypes from closely related individuals. Within a single population, recombination can be tested for using non–parametric approaches such as permutation tests based on summary statistics like the correlation of linkage disequilibrium with distance [50, 63, 3]. Linkage disequilibrium is typically measured using the statistics r^2 and $|D'|$ [39, 24].

Recently, coalescent [34] methods have been developed that can specifically detect [4, 48] or characterize the rate of recombination [15, 23, 36, 53, 76, 11, 26, 48] for a set of samples within a single population. Recombination can either be modeled under a basic crossing over model [25] or a more complex model of gene conversion [80]. Only a few methods [36, 11, 48] relax the infinite sites model [33] under which a site can undergo at most a single mutation. Relaxing the infinite sites model is important for many bacterial and viral data sets, since under the infinite sites model, high levels of recurrent mutation can cause patterns consistent with recombination [48].

The basic coalescent operates under several assumptions that include constant population size, no selection, random mating and no population structure [21]. Whereas these assumptions can be relaxed using additional parameters such as a term for population growth [67], these additional parameters are presently not accounted for in current methods that characterize and detect recombination [11, 36, 48]. Importantly, the influence of population structure and demographic history may adversely affect the ability of coalescent methods to correctly infer the rate of recombination [48, 18].

The myriad of methods available to detect, characterize and find recombinant sequences is somewhat bewildering. Traditionally, general approaches have been used for recombination analysis between distantly related genotypes, whereas population genetic based approaches have

been used for recombination analysis between closely related genotypes. However, in many cases the line between the approaches is blurred, and both approaches have been used to infer the presence of recombination in bacteria, viral and animal mitochondrial data sets [57, 48, 55].

Often, one of the primary questions for any data analysis is to determine whether recombination is likely to be present within a set of sequences at all [3, 48, 57, 44, 55, 74]. Indeed, there are still open questions with regards to the extent of recombination in animal mitochondrial DNA [44, 55, 74]. Moreover, if the sequences are obtained from closely related, yet distinct, organisms or from many different populations, it is inappropriate to analyze the sequences in a framework that assumes a single population, such as linkage disequilibrium or coalescent approaches [74]. But determining whether recombination has occurred in such circumstances is an important question, that cannot be easily answered in a parametric framework. A robust non-parametric test for recombination can help distinguish between the presence and absence of recombination in such cases.

Testing for recombination can statistically validate visual evidence of recombination obtained using, for instance, phylogenetic network approaches (e.g. [28]) or independently verify the presence of recombination if a positive estimate of the rate of recombination is inferred (e.g. [48]). Moreover, it is often difficult to distinguish between rate heterogeneity and recombination in many circumstances [14, 45] and thus regions that exhibit phylogenetic inconsistencies can be individually tested for recombination. Additionally,

testing for recombination can be used as a prior probability for the presence of recombination when inferring the points at which infrequent recombination may have occurred [49]. In this sense, testing for recombination can be used in conjunction with other methods.

Ideally, a single test could correctly determine whether recombination is present within any given set of aligned sequences, regardless of population history, demographic history, recombination rate or mutation rate. Preferably, such a test would also minimize the production of false positives. Here we develop a new test that is powerful under many of these different situations and produces few false positives. Through simulation and empirical data analysis we characterize the performance of our test under various rates of recombination, rates of mutation, demographic histories and sample sizes. We also show through simulation that a simple model of substitution rate autocorrelation (consistent with mutational 'hot spots') gives rise to a signal similar to recombination for two different general tests, Max χ^2 and NSS, but not for our method.

4.4 Methods

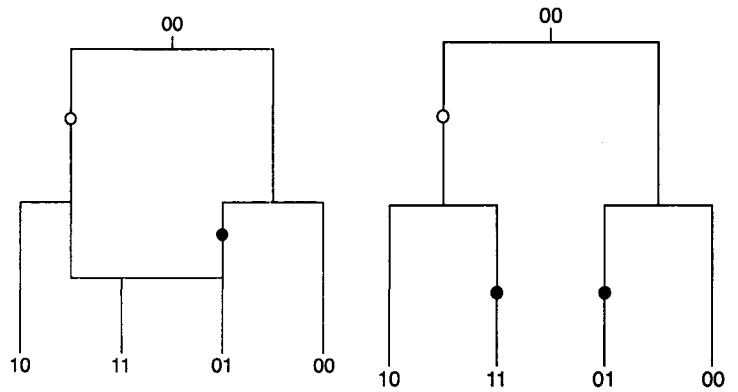
Tests for recombination based on the principle of compatibility have proved to be among the most powerful [4, 58, 79, 57]. The traditional binary notion of compatibility [38] is well suited for sites with at most two alleles, but can be directly extended into a broader notion [54] that we term here as *refined incompatibility.* We then develop a new statistic to test for

recombination, the Φ_w (or Phi) statistic, that uses this notion of refined incompatibility.

4.4.1 Compatibility and Incompatibility

It is not obvious how to determine the genealogical history of a single site. As such, the pattern of mutation present at multiple sites must be used to infer the genealogy of the sample as a whole. One possibility is to use the observed patterns at pairs of sites, in particular the notion of *compatibility* [38] or 'four-gametes' test [27]. Two sites i and j are *compatible* if and only if there is a genealogical history which can be inferred parsimoniously that does not involve any recurrent or convergent mutations (known as *homoplasies* as in Figure 4–1(b)). If the two sites are not compatible, they are termed *incompatible*. Under an infinite sites model [33] of sequence evolution, the possibility of a homoplasy does not exist, and so incompatibility for a pair of sites implies at least one recombination event must have occurred, as Figure 4–1(a). This can be used to estimate the minimum number of recombination events present in the sample as a whole [27, 51, 69]. Testing for compatibility can be accomplished by checking if all four combinations of $\{00, 01, 10, 11\}$ are present among the sequences [38].

The traditional, binary notion of either compatibility or incompatibility treats a single homoplasy the same as many homoplasies. That is, although in some situations more than one homoplasy can be parsimoniously inferred for a pair of sites [7, 54], this information is disregarded. Consider two sites i and j, with $|\chi_i|$ and $|\chi_j|$ representing the number of observed states (alleles)



(a) Two incompatible sites explained by a recombination event

(b) Two incompatible sites explained by a convergent mutation

FIGURE 4–1: The dual nature of incompatibility. The figure shows two possible histories for a pair of incompatible sites. Mutations in the first site are indicated by open circles and mutations in the second site are indicated by solid circles. In order to explain the incompatibility between the pair of sites either a recombination event must be invoked or a homoplasy must have occurred in the history of one of the sites.

at each site. Let $l(\chi_i, \chi_j)$ denote the minimum number of mutations required by *any tree* used to represent the genealogical history of both sites. Thus $l(\chi_i, \chi_j)$ represents the maximum parsimony score for these two characters over all trees. Note that $l(\chi_i, \chi_j) \geq (|\chi_i| - 1) + (|\chi_j| - 1)$ as each state (except the ancestral state) must arise at least once in the tree. Define the *refined incompatibility* score of sites i and j as:

$$i(\chi_i, \chi_j) = l(\chi_i, \chi_j) - (|\chi_i| - 1) - (|\chi_j| - 1)$$

The refined incompatibility score relates to the traditional notion of compatibility in the following way: two sites are compatible if and only if $i(\chi_i, \chi_j) = 0$; if $i(\chi_i, \chi_j) > 0$ the two sites are incompatible. There are also

two interpretations of this refined incompatibility score: in the absence of recombination, this score represents the minimum number of homoplasies that have occurred in the history of the samples for these two sites [54]; in the absence of recurrent or convergent mutations, this score represents the minimum number of recombinations that have occurred between the two sites [5]. This latter result depends on viewing recombinations as unrooted subtree-prune and regraft operations (see [21]). Importantly, this score can be calculated quickly (linear time in the number of sequences [6] which allows alignments with large numbers of sequences to be evaluated rapidly.

A *parsimony informative* site has at least two different alleles that are represented by at least two different sequences each (there must be at least four sequences at a site for the site to be parsimony informative) [12]. A *compatibility matrix* [68, 30] is traditionally used to represent compatibility between all pairs of parsimony informative sites. This matrix can also easily be extended into a *refined incompatibility matrix* by setting each entry (i, j) equal to the refined incompatibility score between any two sites i and j.

Sites that have the same history will tend to be more compatible than sites that have different histories [68, 30, 10]. One way to measure the extent of 'clustering' in the matrix is to consider the proportion of neighboring cells in the matrix that are either compatible or incompatible. The resulting statistic is termed the 'Neighbour Similarity Score' (NSS) and has been used as a powerful test for recombination [30, 4, 58, 79, 57]. However, simulations

suggest that the NSS produces an excess of 'false positives' in certain situations (see Results) and so we have developed an alternative statistic.

4.4.2 Test Statistic (Φ_w)

The degree of genealogical correlation between neighboring sites is negatively correlated with the rate of recombination [27]. In the case of finite levels of recombination, the genealogical correlation of sites is partially reflected by a tendency of closely linked sites to have greater compatibility than distant sites [17, 29].

In order to measure the similarity between closely linked sites, we propose calculating a new statistic, termed the 'Pairwise Homoplasy Index'. The idea is to calculate the mean refined incompatibility score from nearby sites by using the first k off-diagonal rows of a refined incompatibility matrix (see Figure 4–2). Let w denote a fixed width (measured in bases) and choose k so that it is proportional to w. Specifically, let q denote the proportion of parsimony informative sites within the alignment and set $k = wq$. The statistic thus measures the mean refined incompatibility score of sites up to (approximately) w bases apart. We can now formally define the 'Pairwise Homoplasy Index' (Φ or PHI) statistic as:

$$\Phi_w = \frac{2}{k(2n - k - 1)} \sum_{j=1}^{k} \sum_{i=1}^{n-j} i(\chi_i, \chi_{i+j})$$

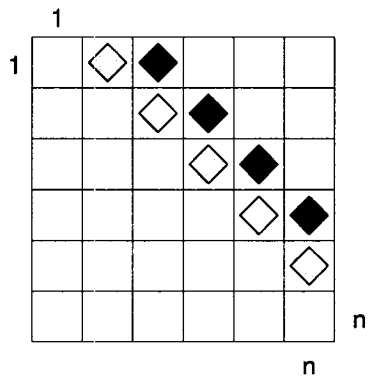


FIGURE 4–2: The entries marked with a diamond in the refined incompatibility matrix represent the cells used to calculate Phi (or Φ_w). The lightly shaded cells contain the refined incompatibility score of informative site i with informative site $i + 1$. The darkly shaded cells contain the refined incompatibility score of informative site i with informative site $i + 2$. In this example sites up to two informative bases apart are used to calculate Φ_w.

The term 'Pairwise Homoplasy Index' refers to the fact that the refined incompatibility score can be interpreted as the minimum number of convergent or recurrent mutations (homoplasies) necessarily present on any tree describing the history of any two sites i and j. The term $k(2n - k - 1))/2$ is a normalizing factor.

Clearly w should be somewhat less than the total number of sites but large enough that a number of comparisons are made. For all simulated and empirical analysis w was set to 100 and k chosen according to the above formula. Other choices of w were also considered ($w = 50$ and $w = 150$), but simulations (across different sequence lengths) suggested that $w = 100$ was slightly better than the other two choices (results not shown).

4.4.3 Significance

Significance of the observed Φ_w statistic can be obtained by using a permutation test. Under the null hypothesis of no recombination, the genealogical correlation of adjacent sites is invariant to permutations of the sites as all sites have the same history. But in the case of finite levels of recombination, the order of the sites is important, as distant sites will tend to have less genealogical correlation than adjacent sites. Let $\hat{z}$ denote the observed value of the Φ_w statistic on the original alignment and let Z_0 denote the value of the Φ_w statistic for a random permutation of the sites. Hence Z_0 is distributed according to the null hypothesis of no recombination. To determine the significance of the observed value $\hat{z}$, a Monte-Carlo p-value can be directly estimated by permuting the alignment many times, and counting the proportion of times the Φ_w statistic on a permuted alignment is less than or equal to $\hat{z}$. However, computation of p-values based on permutations of the alignment is time consuming. One way to circumvent this problem is to determine the distribution of the test statistic under permutations of the alignment. The expectation ($\mathrm{E}_0(\Phi_w) = \mu'$) and variance ($\mathrm{Var}_0(\Phi_w) = \sigma^2$) of Φ_w can be calculated analytically (see Appendix A for details). Moreover, initial simulations indicated that the distribution of Φ_w under permutations of the alignment is approximately normal (results not shown). Using these assumptions, the

value of $\Pr(Z_0 \leq \hat{z})$ can be calculated as:

$$\Pr(Z_0 \leq \hat{z}) = \int_{-\infty}^{\hat{z}} n(\tau|\mu', \sigma^2) d\tau$$

where $n(\tau|\mu', \sigma^2)$ denotes a normal probability distribution function with mean μ' and variance σ^2. This alternative to the permutation test has the advantage that it can be obtained quickly and gives a more precise p-value under an assumption of normality.

The normality of the distribution of the test statistic can be explained by noting that for a large refined incompatibility matrix, calculating the Φ_w statistic amounts to taking the mean of a small sample of values from the matrix. The simplest version of the Central Limit Theorem then suggests that taking the mean of a small sample within a 'large' matrix has a limiting normal distribution, if the terms are independent and identically distributed [8]. However, in this case the Central Limit Theorem provides a guide rather than a formal equivalence.

For every data set examined (both simulated and empirical) the significance of the observed Φ_w statistic was calculated using both the permutation test directly as well as the normal alternative. The p-values obtained by using the permutation test are written as $\mathrm{P_P}(\Phi_w)$ whereas the p-values obtained by using the normal alternative are written as $\mathrm{P_N}(\Phi_w)$.

4.4.4 Simulation Study

We repeated many of the same simulations that had been performed in other studies [58, 79] but expanded the parameter search space and

considered the Φ_w statistic as well as additional tests. The protocol followed was based on simulations from the neutral coalescent model [34] with recombination [25].

The coalescent model provides a natural foundation for simulation [9, 4, 58, 79]. Simulations were almost all conducted using the program Treevolve [13]. For very high rates of recombination ($\rho = 128$), simulations were performed using the program Hudson [65, 64] since the program Treevolve did not run at such high rates of recombination. Mutations were added according to a Jukes–Cantor model [32]. Other methods of sequence evolution were also examined, including the addition of extreme rate heterogeneity ($\alpha = 0.1$) which resulted in a moderate decrease in power for all methods (results not shown). For each parameter setting, 1000 replicate data sets were created, with each replicate consisting of an alignment of length 1000 (see Appendix B for further details). Significance was set at the 0.05 level.

In addition to the Φ_w statistic, four of the best non-parametric tests were computed for each parameter setting, namely the Max χ^2 [43] statistic, the 'Neighbour Similarity Score' (NSS) [30], and two measures of correlation of linkage disequilibrium (r^2 and $|D'|$) with distance [39, 24, 50, 63]. Furthermore, results obtained from a coalescent based likelihood permutation test (LPT) from LDHat [48] are reported as well. The Max χ^2 statistic has been found to be the best general test for detecting recombination in a recent empirical study [57], and the NSS statistic has been found to be very

efficient as well [4, 58, 79, 57]. Correlation of linkage disequilibrium with distance using r^2 has been found to be the strongest non-parametric approach for detecting recombination within populations [48]. Recently, the likelihood permutation test was introduced as a powerful alternative to methods based on linkage disequilibrium [48]. For the Max χ^2 statistic a fixed window size of the number of polymorphic sites divided by 1.5 was used following a previously described protocol [58, 57]. For both correlation of r^2 and D' with distance, only sites with two alleles segregating and minor allele frequencies of at least 0.1 were used, as this approach tends to maximize power [78, 48]. For the likelihood permutation test, precomputed likelihood files were used based on 101 grid points with a value of θ per site either equal to 0.001 or 0.1. For each replicate, if the expected mean sequence diversity was less than 10%, then a likelihood file with a θ per site value of 0.001 was used, otherwise a likelihood file with a θ per site value of 0.1 was used (under a constant size population the expected mean sequence diversity of 10% corresponds to an expected value of θ per site of about 0.12). The significance for each of the statistics was obtained using a permutation test. For the power determination, 1000 permutations were performed, whereas for the false positives, 200 permutations were performed.

Power: In order to determine power in the presence of recombination, the recombination rate ρ (under population growth $\rho^{\dagger}$) varied among $0, 1, 2, 4, 8, 16$ and 128, the expected nucleotide diversity p between any two sequences varied among $1\%, 5\%, 10\%, 15\%$, and 25%, and the growth

rate of the population β varied between 0 (constant sized populations) and 5000. The sample size m varied among $5, 10, 15, 25$, and 50. For $\rho = 128$ simulations with $\beta = 5000$ were not performed since this option was not available with the program Hudson. More details explaining the protocol can be found in Appendix B and elsewhere [79].

False Positives: Substitution rate heterogeneity across sites on a genealogy was modeled here using a Γ distribution [75, 81]. In this case, the substitution rate at each site i, Z_i, is drawn from a Γ distribution with shape parameter α and scale parameter $1/\alpha$ [81].

Auto–correlation among substitution rates was modeled assuming Markov–dependence among rates [82]. To achieve this, two random variables Y_i and Y_{i+1} were drawn from a bivariate normal distribution with correlation ρ_N and transformed into two marginally distributed gamma random variables Z_i and Z_{i+1} with correlation ρ_G [82]. Using the bivariate normal distribution of Y_i and Y_{i+1} (including correlation ρ_N), the probability distribution function of random variable Y_{i+1} was obtained conditional on the random variable Y_i, allowing Markov dependent substitution rates to be drawn. The substitution rate Z_i and Z_{i+1} then represent draws from a bivariate Γ distribution with correlation ρ_G. The value of ρ_G is positively correlated with the value ρ_N but not identical [82].

Data sets were simulated using a modified version of Treevolve [13] with a number of the sampling functions taken from PAML [83]. The correlation parameter ρ_N varied among 0 (no correlation), $0.3, 0.6$ and

0.9, the expected nucleotide diversity p between any two sequences varied among $1\%, 5\%, 10\%, 15\%$, and 25%, the value of α for the Γ distribution varied among $0.1, 1.0$ and ∞, and the growth rate of the population β varied between 0 (constant sized populations) and 5000. The sample size m varied among $5, 10, 15, 25$, and 50.

4.4.5 Empirical Data

A number of population and species level data sets were examined. The presence of recombination in each of these data sets was either debated, unknown or suspected. The rate of recombination in these data sets ranged from rare to very frequent. In general, data sets with at least a few hundred sites were chosen.

Tests for recombination were performed using the Φ_w statistic as well as the Max χ^2 statistic [43] and the NSS statistic [30]. As in the simulation studies, w was set to 100 for all analysis. One thousand permutations were performed to obtain significance. Additional results are reported for the population level data sets, using permutation tests based on r^2 and $|D'|$ [39, 24, 50, 63] as well as a coalescent based likelihood permutation test (LPT) with LDhat [48]. Furthermore, an estimate of the rate of recombination was also obtained in LDHat using a model of crossing-over rather than gene conversion. The maximum value of ρ was set to 100 and 100 grid points were used in LDHat. The value of Tajima's D statistic is also reported, as it can be an indicator of population growth or selective pressure [73]. Table 4–1 summarizes the data sets used. The data sets

include sequences from bacteria, viruses and fungi. Two of the data sets were from animal mitochondrial DNA (mtDNA).

TABLE 4–1: Summary of empirical data sets

Data Set	Type	Number of Sequences	Number of Sites	Informative Sites	Observed Diversity[a]	Tajima's D[b]	Reference
C. albicans	Fungi	45	2553	58	0.7%	0.936	[1]
Rana	mtDNA	8	1143	257	14.8%	-	[71]
C. ruminantium	Bacteria	14	870	186	10.5%	0.384	[31]
H. pylori	Bacteria	33	472	53	3.8%	−0.531	[70]
Boletales	Fungi	31	639	265	17.1%	-	[35]
Norovirus	Virus	25	1617	103	2.2%	-1.482	[60]
Apodemus	mtDNA	10	1140	275	14.7%	-	[41]
N. *Wolbachia*	Bacteria	10	444	98	13.0%	0.899	[31]

[a] Mean proportion of sites that differ between any two sequences.
[b] Calculated on sites with only two alleles segregating.

For the Boletales data set additional analysis was performed by first estimating a neighbor-joining tree [61] using PAUP* [72]. Branch lengths for the tree, a transition/transversion ratio, codon frequencies, a value of α for the substitution rate heterogeneity [81], as well as the degree of substitution rate autocorrelation (estimated using the auto–discrete gamma model) [82], were then estimated using a codon model in PAML [83]. A parametric bootstrap of 1000 replicates was then performed under the estimated parameters using a modified version of PAML that allowed autocorrelated substitution rates. For each replicate, a test for recombination was performed using the Max χ^2 statistic, the NSS statistic and the Φ_w statistic (with 1000 permutations). Significance was set at 0.05.

4.5 Results and Discussion

4.5.1 Analytical Calculation of p-values

Table 4–2 shows the proportion of times that recombination was inferred using Φ_w, when the rate of recombination ρ was set to 0 and there was no population growth ($\beta = 0$). Since the significance level was set to 0.05, the Φ_w test is too conservative when the mean sequence diversity is about 1% or when there are few samples (e.g. $m = 5$). This is partly due to the fact that there are very few informative sites or incompatibilities produced in these situations (results not shown). Table 4–2 also indicates that when the sequence diversity and sample size is small, obtaining significance using the permutation test ($P_P(\Phi_w)$) is even more conservative

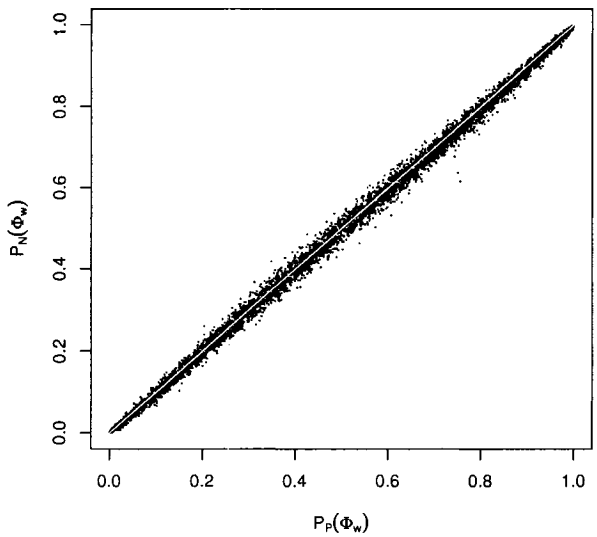


FIGURE 4–3: Comparison of p-values obtained using the permutation test (horizontal axis) to analytical p-values (vertical axis) when $\rho = 0$ and $\beta = 0$. Points with less than 15 samples and less than 10% sequence divergence are not shown (see Table 4–2).

than obtaining significance using the normal distribution ($P_N(\Phi_w)$). On the other hand, Figure 4–3 shows that both methods for obtaining significance give very similar answers for higher amounts of sequence diversity (at least 10%), with at least 15 samples. These results suggest that it is sufficient to obtain significance for Φ_w using the normal distribution. For all subsequent simulations, the results quickly obtained with the Φ_w statistic using the normal distribution are reported.

TABLE 4–2: Percentage of time recombination inferred using Φ_w when $\rho = 0$ and $\beta = 0$ (without mutation rate correlation or substitution rate heterogeneity).

	Diversity									
m	1%		5%		10%		15%		25%	
5	0.4	0.4	1.6	0.9	3.6	1.7	4.2	2.4	5.1	3.7
10	0.1	0.0	3.1	1.5	4.6	3.5	3.9	3.2	4.7	4.0
15	0.2	0.0	5.5	3.8	5.7	4.7	5.4	4.5	4.0	3.8
25	0.3	0.2	4.6	2.9	4.8	4.3	4.5	3.8	4.5	4.1
50	0.8	0.1	5.9	4.5	4.1	3.8	5.7	5.6	5.7	5.3

The columns for each parameter pair represent $P_N(\Phi_w)$ and $P_P(\Phi_w)$ respectively.

4.5.2 Time

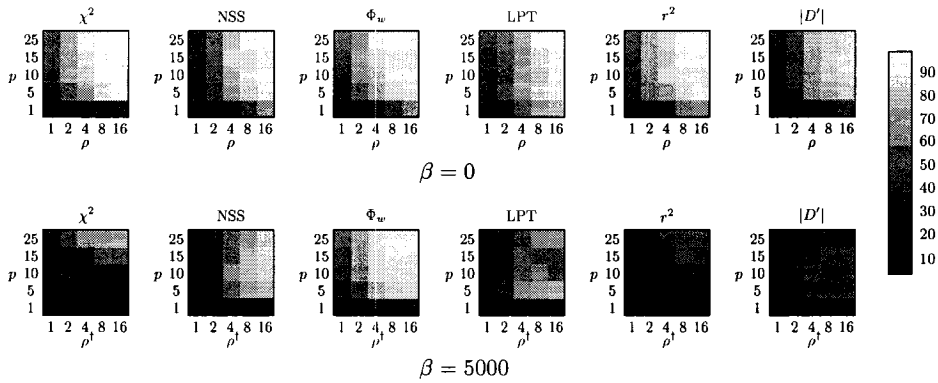
The time to calculate Φ_w is much faster than other population genetic methods especially for moderate numbers of sites and sequences. For instance, several simulated alignments of 25 samples with 5000 sites with moderate sequence diversity (10%), corresponding to viral genomic samples, were analyzed on a Mac G4 desktop computer. The time taken to analyze each alignment was about twenty seconds using Φ_w without the permutation test, thirty seconds using Φ_w with the permutation test, seven minutes with the linkage disequilibrium methods (using LDHat), and 8 hours using the likelihood permutation test of LDHat (using a pre-computed likelihood file). For longer alignments however, the permutation test becomes impractical even for Φ_w and in these cases analytical p–values are the only way to practically test for recombination. It is worth noting that since the power to detect recombination increases as a function of sequence length [79], this

constitutes an important advantage for the Φ_w test, since faint recombinant signals may be only detectable using very long sequences.

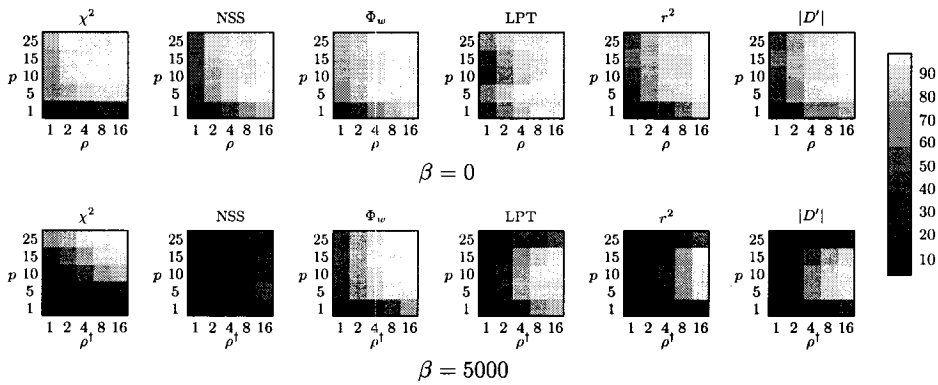
4.5.3 Power

Figure 4–4 shows the power to detect recombination for Φ_w, Max χ^2, NSS, the likelihood permutation test (LPT) in LDHat and two measures of correlation of linkage disequilibrium with distance (r^2 and $|D'|$), when the rate of recombination ρ is greater than zero, for two different sample sizes ($m = 10$ and $m = 50$). Two principal types of genealogies were created: with and without population growth. If there is population growth, the genealogies created will be more starlike with long branches at the leaves [16, 79]. If there is no population growth, there are short branches at the tip but long branches at the root. When genealogies are more starlike, recurrent mutations will tend to mask the initial recombination, and the recombination events are best considered 'ancestral'.

The top rows of Figure 4–4(a) and 4–4(b) show that without population growth ($\beta = 0$), all six methods performed similarly, although overall Φ_w is the most powerful method with a large number of samples. Without population growth, the power to detect recombination of all six methods generally increases as a function of both sequence diversity and the rate of recombination, similar to earlier observations [58, 79]. A notable exception is the likelihood permutation test (LPT) for which there is a slight decline in power when the mean sequence diversity reaches 10%. At this point, a likelihood file with a value of θ per site of 0.1 was used rather than



(a) $m = 10$ samples



(b) $m = 50$ samples

FIGURE 4–4: Power to detect recombination for a) $m = 10$ and b) $m = 50$ samples for six different methods with (bottom row of each subfigure) and without (top row of each subfigure) population growth. The horizontal axis varies the rate of recombination whereas the vertical axis varies the amount of sequence diversity. Each cell represents the outcome of 1000 replicates with lighter cells indicating increased power. The value $\rho^\dagger$ refers to the value of ρ used to give the same expected number of recombinations under population growth.

a likelihood file with a value of θ per site of 0.001. However, when the sequence diversity reach 10%, the expected value of θ per site is about 0.12 suggesting that a value of θ per site of 0.1 is a better choice. Nonetheless, more power may be obtained by using a gross underestimate of θ, although previous work has demonstrated a relative insensitivity of the LPT to a specific estimate of θ [48].

The top rows of Figure 4–4(a) and Figure 4–4(b) suggest that the Φ_w method performs similarly to the linkage disequilibrium approaches when there is very little sequence diversity (e.g. $p = 1\%$), despite the fact that the test is too conservative in these circumstances (Table 4–2). For very little sequence diversity (i.e. $p = 1\%$), the coalescent based method LPT is the most powerful method in constant sized populations, but has about the same power as Φ_w for growing populations. However, the results suggest that all methods may underestimate the presence of recombination if few sequences are present with very little divergence, especially in an expanding population (or 'starlike' genealogy).

By comparing the bottom rows of Figure 4–4(a) and 4–4(b) to the top rows of Figure 4–4(a) and 4–4(b) it is evident that detecting the presence recombination under population growth ($\beta = 5000$) is a more difficult task than detecting the presence of recombination without population growth ($\beta = 0$). Out of all six methods, the bottom rows of Figure 4–4(a) and 4–4(b) suggest that Φ_w is much better at detecting recombination under population growth than either Max χ^2, NSS, the coalescent based LPT or

the linkage disequilibrium approaches. For the coalescent based LPT, it is worth noting that population growth could be incorporated in the method in the future, possibly increasing power. The decline of linkage disequilibrium in expanding populations using r^2 is consistent with previous observations [66, 47], but the results suggest that the performance of the $|D'|$ statistic is similar. The results for the Φ_w test suggest that subsequent mutations do not 'mask' the recombinant signal for this method. Interestingly, this is similar behavior to the RECPARS method [20, 79] and may be of particular importance when trying to determine ancestral recombination between diverged genotypes. The results also suggest that the Φ_w statistic can be used to distinguish between starlike genealogies due to population growth and starlike genealogies to recombination [65].

A comparison of the top row of Figure 4–4(a) to the top row of Figure 4–4(b) reveals that an increase in sample size from $m = 10$ to $m = 50$ causes an increase in the ability of all six methods to infer recombination when there is no population growth ($\beta = 0$). For population growth (the bottom rows of both Figure 4–4(a) and 4–4(b)), the power to detect recombination for the NSS statistic for actually decreases sharply from $m = 10$ to $m = 50$. But for the other five tests, the power to detect recombination generally increases when moving from $m = 10$ to $m = 50$ even under population growth. These results expand upon some previous observations [79].

Under a neutral coalescent model with recombination, it is possible to use a likelihood ratio test to determine whether the hypothesis of no

recombination ($\rho = 0$) should be rejected at a given significance level [36, 4]. However, even when data is simulated according to the neutral coalescent with low levels of recombination, the hypothesis $\rho = 0$ is only rejected a limited proportion of the time [4]. However, such a simulation represents an ideal situation, where the likelihood ratio test is guaranteed to be the most powerful [4] and the model used to infer ρ is identical to the model used to generate samples. This suggests it might be difficult for any test to correctly infer the presence recombination for very low recombination rates. Additionally, a theoretical analysis shows that generating small sets of samples using a low rate of recombination only produces a limited number of incompatibilities [79]. It is thus possible that full likelihood approaches [36, 11] or a phylogenetic network [28] approach could be particularly useful to determine whether there is any possibility of recombination when only a weak recombinant signal exists.

Table 4–3 demonstrates that Φ_w can detect recombination even under extremely high recombination rates ($\rho = 128$). Except for low sequence diversity ($p = 1\%$), the presence of recombination is correctly inferred each time. But even for low sequence diversity, the presence of recombination can be inferred nearly every time by increasing the sample size from $m = 10$ to $m = 50$.

It is worth noting that the Φ_w statistic can also be calculated without the refined incompatibility score, but using only the traditional notion compatibility. For cases without population growth ($\beta = 0$), the results are

TABLE 4–3: Power to detect recombination using Φ_w with a high rate of recombination $\rho = 128$

Diversity	Number of Samples	
	$m = 10$	$m = 50$
1%	68%	99%
5%	100%	100%
10%	100%	100%
15%	100%	100%
25%	100%	100%

almost identical (results not shown). On the other hand, with population growth ($\beta = 5000$), there is an increase in power using the refined incompatibility score when the number of samples is large (e.g. $m = 50$) and there is some recurrent mutation. For a rate of recombination of $\rho = 1$, a sample size of 50, and exponential growth, the gain in power using the refined incompatibility score rather than the compatibility score was $2\%, 5\%$ and 12% for mean pairwise sequence divergences of $10\%, 15\%$ and 25% respectively. Similar results are obtained for $\rho = 2$ but not for higher rates of recombination (results not shown). This suggests that the refined incompatibility score is a useful extension to the traditional notion of compatibility especially for large sample sizes with sites that experience recurrent mutations.

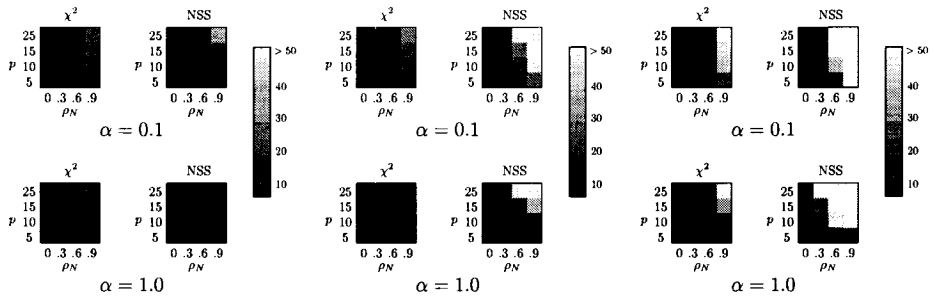
For no population growth, the Φ_w test, and the linkage disequilibrium approaches perform similarly, although Φ_w is more powerful for a large number of samples. However Φ_w is applicable even if the samples are from different species or different populations, whereas the linkage disequilibrium and coalescent approaches are not [74]. Under population growth however

($\beta = 5000$), only Φ_w continues to consistently infer the presence of recombination as the power of the other five methods suffer sharp declines. This suggests out of all six methods, Φ_w has the greatest flexibility in detecting recombination in the different circumstances studied.

4.5.4 False positives

Of particular concern for any test for recombination is the effect of confounding processes such as substitution rate heterogeneity and autocorrelated substitution rates. Autocorrelation of substitution rates imply that the rate of substitution of one site is not independent of the rate of substitution of a neighboring site and can create 'mutational hot–spots' within a sequence. This can potentially create the same patterns as recombination.

Figure 4–5 shows the proportion of false positives for Max χ^2 and NSS when there is no recombination ($\rho = 0$) but 'mosaic' sequences are artificially induced by using a range of autocorrelated substitution rates. Figure 4–5 shows that both Max χ^2 and NSS falsely infer the presence of recombination more than 50% of the time in certain cases. The results for the linkage disequilibrium, likelihood permutation test and Φ_w are omitted from Figure 4–5 since these methods did not falsely infer recombination more than 7% of the time, although Table 4–4 shows this information for Φ_w. Table 4–4 shows the Φ_w statistic did not infer recombination more than 6% of the time when recombination was falsely inferred more than 50% of the time using both Max χ^2 and NSS. Although the global model



(a) $m = 10$ samples ($\beta = 5000$) (b) $m = 50$ samples ($\beta = 0$) (c) $m = 50$ samples ($\beta = 5000$)

FIGURE 4–5: Percentage of false positives for a) $m = 10$ samples (with $\beta = 5000$), b) $m = 50$ samples (with $\beta = 0$), and c) $m = 50$ samples (with $\beta = 5000$), for Max χ^2 and NSS, with extreme rate heterogeneity (top row) and moderate rate heterogeneity (bottom row). The horizontal axis varies the substitution rate correlation whereas the vertical axis varies the amount of sequence diversity. Each cell represents the outcome of 1000 replicates with lighter cells indicating a higher percentage of false positives. The results for Φ_w, r^2 and $|D'|$ are omitted since these approaches did not falsely infer falsely recombination more than 7% of the time for any of the conditions, but Table 4–4 shows a number of these results for Φ_w.

of substitution rate autocorrelation employed by this study is quite simple since it ignores codon positions and substitution rate correlation within local patterns of substitution [46], it nonetheless provides a guide as to the effect of autocorrelated substitution rates.

The problem of false positives in NSS and Max χ^2 is most severe for large sample sizes (e.g. $m = 50$), both under constant sized populations (Figure 4–5(b)) and population growth (Figure 4–5(c)) . Although the problem is in general greater for higher substitution heterogeneity (the top rows of Figure 4–5(a), 4–5(b), and 4–5(c)) it is also a problem with lower

TABLE 4–4: Proportion of times recombination is falsely inferred using Φ_w with substitution rate heterogeneity $\alpha = 0.1$, mutation rate correlation and sample size $m = 50$.

	Mutation Rate Correlation							
Diversity	0		0.3		0.6		0.9	
1%	2.0	3.6	2.5	3.6	2.6	3.9	1.1	3.8
5%	4.9	4.7	5.8	4.5	4.7	3.3	3.0	1.0
10%	4.1	5.6	4.7	4.6	4.8	3.0	1.8	1.5
15%	4.9	4.0	4.5	4.7	3.8	4.5	2.9	1.8
25%	5.3	4.0	3.7	3.5	4.1	3.9	3.4	2.1

The columns for each parameter pair represent the outcomes for $\beta = 0$ and $\beta = 5000$ respectively.

substitution rate heterogeneity (the bottom rows of Figure 4–5(a), 4–5(b), and 4–5(c)).

The level of false positives of both NSS and Max χ^2 suggests caution in interpreting evidence for recombination, especially when autocorrelated rates are an issue. For instance, inferring the presence of recombination in mitochondrial DNA should be done cautiously as substitution rate correlation is known [82, 52].

The results using Φ_w contrast strongly with the results using the NSS (which is also compatibility based). This is likely due to the difference in the statistics themselves. The Φ_w statistic uses compatibility between closely linked sites directly whereas the NSS statistic measures 'clustering' within a compatibility matrix. As the clustering can be caused by substitution rate correlation, and not only recombination, this might explain the difference between the two statistics. For Max χ^2 the problem is possibly due to pairs

of sequences that differ greatly on one side of a site (due to high mutation) but share a great degree of similarity on the other side of a site (due to low mutation). Local 'bursts' of mutation [46] likely exacerbate the problem, especially for linkage disequilibrium approaches that are based on allele frequencies at different sites.

4.5.5 Empirical Data

The general information concerning the empirical data sets is summarized in Table 4–1. Tables 4–5 and 4–6 show the results of tests for recombination on all the empirical data sets. In addition to the results obtained using the Φ_w statistic, results using Max χ^2 [43], NSS [30], correlation of r^2 and $|D'|$ with distance [39, 24] and a likelihood permutation test (LPT) [48] are shown. The estimates of ρ for the population level data sets were obtained using LDHat [48]. Tests for recombination within populations (i.e. r^2, $|D'|$ and LPT) were not applied to data sets that contained individuals from different species.

4.5.6 Recombinant Examples

Table 4–5 shows that the null hypothesis of no recombination is rejected by all tests for most of the suspected recombinant data sets, including the *Candida* example that had very little sequence diversity (0.7%). Whereas a lack of sequence diversity in the simulations made recombination harder to detect, this may be partially overcome by using longer alignments, such as for the *Candida* example which had 2553 sites.

Interestingly, the null hypothesis of no recombination was not universally rejected for two of the bacterial data sets: *Cowdria* and *H. pylori.* For these two bacterial examples, evidence for recombination was found using the Φ_w statistic as well as the coalescent based likelihood permutation test. However, recombination was only detected in the *Cowdria* example using the correlation of distance with r^2 after sites with minor alleles were removed. Moreover, in the *H. pylori* data set neither NSS nor Max χ^2 found significant evidence for recombination. This could be due to the high suspected rate of recombination in the *H. pylori* example, which has conditions approaching linkage equilibrium [70]. The linkage disequilibrium methods seem to be highly sensitive to sites with low allele frequencies and consistent results are only obtained after the removal of these sites.

TABLE 4–5: Analysis of suspected recombinant data sets

Data Set	ρ[a]	Φ_w[b,c]	χ^2	NSS	r^2[a,d]	$\lvert D' \rvert$[a,d]	LPT[a,d,e]
Candida	16	2.4×10^{-15}* (.000*)	.000*	.000*	.000* (.000*)	.122 (.001)	.000* (.000*)
Rana	-	5.5×10^{-31}* (.000*)	.000*	.000*	-	-	
Cowdria	17	3.8×10^{-05}* (.000*)	.041*	.001*	.167 (.039*)	.043* (.029*)	.000* (.001*)
H. pylori	≥ 100	9.3×10^{-03}* (.004*)	.158	.330	.125 (.000*)	.536 (.003*)	.000* (.000*)

* $P < .05$

[a] Calculated on sites with only two alleles segregating with LDHat.

[b] Each pair shows p-values calculated analytically and using a permutation test respectively.

[c] w was set to 100 for all tests.

[d] Terms in parenthesis show results on sites with minor allele frequencies > 0.1.

[e] Denotes the value of a likelihood permutation test calculated in LDHat.

4.5.7 Possibly Recombinant Examples

The results obtained from the data sets for which the status of recombination is debated are quite interesting (Table 4–6). For the Norovirus example, evidence of recombination is found using Φ_w, Max χ^2 and the LPT. There is some evidence of recombination found with r^2, but after sites with minor allele frequencies less than 0.1 are removed no further evidence is found by the linkage disequilibrium methods. Since the samples came from a number of different cities, it could be that evidence of recent recombination is weakened by removing these sites. However, the LPT finds evidence of recombination regardless of whether or not these sites are removed.

TABLE 4–6: Analysis of possibly recombinant data sets

Data Set	ρ^{a}	$\Phi_w{}^{b,c}$	χ^2	NSS	$r^{2a,d}$	$\lvert D' \rvert^{a,d}$	$\text{LPT}^{a,d,e}$
Norovirus	23 (21)	.002* (.003*)	.025*	.237	.029* (.574)	.868 (.340)	.022* (.026*)
Apodemus	-	.135 (.151)	.274	.006*	-	-	-
Boletales	-	.934 (.931)	.003*	.000*	-	-	-
Wolbachia	0 (2)	.086 (.103)	.566	.108	.049* (.019*)	.286 (.204)	.709 (.090)

* $P < .05$

[a] Calculated on sites with only two alleles segregating with LDHat.

[b] Each pair shows p-values calculated analytically and using a permutation test respectively.

[c] w was set to 100 for all tests.

[d] Terms in parenthesis show results on sites with minor allele frequencies > 0.1.

[e] Denotes the value of a likelihood permutation test calculated in LDHat.

For the bacterial symbiont nematode *Wolbachia*, there is little prior reason to suspect recombination [31]. Nonetheless, evidence for recombination is found using correlation of r^2 with distance and marginal evidence for recombination is found by using the likelihood permutation test when sites with minor alleles frequencies less than 0.1 are removed. The results obtained using the Φ_w statistic also suggest that there is marginal evidence for recombination with *Wolbachia*. The possible presence of recombination in *Wolbachia* should be tested further using more data.

Recombination in the animal mitochondrial DNA of *Apodemus* was first proposed [37] and then disputed [44]. Tests for recombination using Φ_w and Max χ^2 indicate that there is little evidence for recombination, although the NSS statistic does find evidence for recombination. The evidence for recombination within *Apodemus* using the Max χ^2 test is even weaker here than in previous studies [44], possibly due to the fact that this implementation of the Max χ^2 test uses a 'fixed window size'. Given the high level of false positives of NSS, the results suggest evidence for recombination within *Apodemus* is lacking.

For the fungal Boletales, results using the Φ_w statistic are quite distinct from the results obtained using both the NSS and Max χ^2 statistic. The Φ_w based tests find no evidence for recombination whereas both other tests find strong evidence for recombination. Interestingly, although most other methods for detecting recombination find evidence for recombination

within this data set, Geneconv [62], another powerful sequence-based test for recombination, does not [57].

One possibility for the Boletales data set is that the Φ_w statistic is too conservative and produced a Type II error ('false negative'). A Type II error is the error of not rejecting the null hypothesis when the alternative hypothesis is true. The Boletales data set is a saturated data set with a strong A+T bias [35]. The strong A+T bias results in an estimated transition/transversion ratio of 0.4. Simulation show however, that even under such conditions, there is reason to believe that recombination will still create distinct patterns of compatibility and incompatibility that should be detectable using the Φ_w statistic (results not shown). Moreover, simulations indicate that the Φ_w statistic appears to be more powerful than the NSS statistic (which is also compatibility based) suggesting that a Type II error for the Φ_w statistic, but not for the NSS statistic, is unlikely.

Another possibility for the Boletales example is that both Max χ^2 and the NSS statistic are producing Type I errors ('false positives'), which, according to the simulations, autocorrelated substitution rates might induce. A Type I error is the error of incorrectly rejecting the null hypothesis when it is true. To test this, a parametric bootstrap with 1000 replicates simulating codons (with no recombination) was performed using a substitution rate heterogeneity of 1.31 and global substitution rate correlation $\rho_G = 0.35$ as estimated from the data set. Figure 4–6 shows the distribution of estimated p-values obtained on the 1000 replicates using the Max χ^2 statistic, NSS

statistic and Φ_w statistic. Recombination was inferred 5.7% of the time using the Φ_w statistic, 8.5% of the time with the Max χ^2 statistic and 37.5% of the time using the NSS statistic. Since none of the replicates contained recombination, the p-values for each of the three methods should follow a uniform distribution. Figure 4–6 shows that the parametric bootstrap creates conditions similar to recombination for both Max χ^2 and NSS (a one-sided Kolmogorov-Smirnov test [42] rejects the uniform distribution at a significance level of 10^{-7} for both Max χ^2 and NSS but fails to find any evidence to reject the uniform distribution for Φ_w). Whereas the results for Max χ^2 are less striking than those for NSS, the parametric bootstrap fails to account for local patterns of mutation [22, 46, 48], which are likely to exacerbate the observed bias. These results suggest that there is reason to doubt the validity of the inferences of Max χ^2 and NSS concerning the presence of recombination in the Boletales data set.

4.6 Conclusion

We have presented a simple, powerful test to detect recombination that can be used regardless of sample history. The approach is very general (e.g. does not assume a single population) and aims simply to determine whether there is a recombinant signal present within the sequences. In contrast to two other general tests, Max χ^2 and NSS, our test does not falsely infer the presence of recombination because of mutation rate correlation (which is present in some mitochondrial DNA). Interestingly, our approach performs very well even in the presence of population growth, in contrast

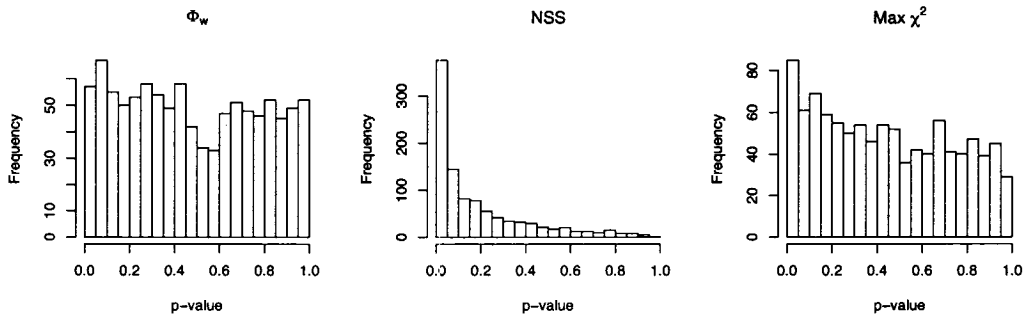


FIGURE 4–6: Distribution of p-values inferred by the Φ_w statistic, the NSS statistic and the Max χ^2 statistic. The results are obtained based on 1000 parametric bootstraps under conditions observed for the Boletales example. None of the replicates contained recombination but the substitution rate autocorrelation was set to $\rho_N = 0.35$ and substitution rate heterogeneity was set to $\alpha = 1.31$.

to methods based on linkage disequilibrium (r^2 and $|D'|$), a coalescent based likelihood permutation test (from LDHat), Max χ^2 and NSS. Our method can be used either by itself, or to validate the visual presence of recombination from a phylogenetic network approach, or to independently verify the presence of recombination if a positive estimate of the rate of recombination is obtained. The approach may be particularly useful in distinguishing recurrent mutation from recombination when assumptions such as a single, randomly mating, and constant sized population are not met. The test can easily be used when many sequences and sites are present because of its computational efficiency and indeed is more powerful in such circumstances. A program implementing our test as well as both Max χ^2 and NSS is available as a standalone program at the following address:

http://www.mcb.mcgill.ca/~trevor. The test is also implemented in SplitsTree 4.2, available at: http://www.splitstree.org.

Acknowledgements

TB would like to thank Kirk and Rachel Bevan, Scott Bunnell, Daniel Huson and Russell Steele, as well as the two anonymous referees for a number of helpful suggestions that greatly improved the manuscript. TB is supported by a National Science Engineering and Research Council (Postgraduate Scholarship B) and Fonds de recherche sur la nature et les technologies (FQRNT grant 2003-NC-81840). DB is supported in part by National Science and Engineering Research Council (NSERC grant 238975-01). HP acknowledges Génome Québec.

References

[1] J. B. Anderson, C. Wickens, M. Khan, L. E. Cowen, N. Federspiel, T. Jones, and L. M. Kohn. Infrequent genetic exchange and recombination in the mitochondrial genome of *Candida albicans*. *Journal of Bacteriology*, 183(3):865–72, 2001.

[2] P. Awadalla. The evolutionary genomics of pathogen recombination. *Nature Reviews Genetics*, 4(1):50–60, 2003.

[3] P. Awadalla, A. Eyre-Walker, and J. Maynard Smith. Linkage disequilibrium and recombination in hominid mitochondrial DNA. *Science*, 286(5449):2524–2525, 1999.

[4] C. J. Brown, E. C. Garner, A. K. Dunker, and P. Joyce. The power to detect recombination using the coalescent. *Molecular Biology and Evolution*, 18(7):1421–4, 2001.

[5] T. C. Bruen and D. Bryant. Maximum parsimony is a consensus method. *submitted*, 2006.

[6] T. C. Bruen and D. Bryant. A subdivision approach to maximum parsimony. *Annals of Combinatorics*, (In press), 2006.

[7] J. H. Camin and R. R. Sokal. A method for deducing branching sequences in phylogeny. *Evolution*, 19(3):311–326, 1965.

[8] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury Press, 2001.

[9] K. A. Crandall and A. R. Templeton. Statistical approaches to detecting recombination. In Keith A. Crandall, editor, *The Evolution of HIV*, pages 153–176. Johns Hopkins University Press, 1999.

[10] G. Drouin, F. Prat, M. Ell, and G. D. Clarke. Detecting and characterizing gene conversions between multigene family members. *Molecular Biology and Evolution*, 16(10):1369–90, 1999.

[11] P. Fearnhead and P. Donnelly. Estimating recombination rates from population genetic data. *Genetics*, 159(3):1299–318, 2001.

[12] J. Felsenstein. *Inferring Phylogenies.* Sinauer Associates, 2004.

[13] N. C. Grassly, P. H. Harvey, and E. C. Holmes. Population dynamics of HIV-1 inferred from gene sequences. *Genetics*, 151(2):427–38, 1999.

[14] N. C. Grassly and E. C. Holmes. A likelihood method for the detection of selection and recombination using nucleotide sequences. *Molecular Biology and Evolution*, 14(3):239–47, 1997.

[15] R. C. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, 3(4):479–502, 1996.

[16] R. C. Griffiths and S. Tavaré. The age of a mutation in a general coalescent tree. *Stochastic Models*, 14:273–295, 1998.

[17] J. Hagenblad and M. Nordborg. Sequence variation and haplotype structure surrounding the flowering time locus FRI in *Arabidopsis thaliana*. *Genetics*, 161(1):289–98, 2002.

[18] D. T. Haydon, A. D. S. Bastos, and P. Awadalla. Low linkage disequilibrium indicative of recombination in foot-and-mouth disease virus gene sequence alignments. *Journal of General Virology*, 85:1095–100, 2004.

[19] J. Hein. Reconstructing evolution of sequences subject to recombination using parsimony. *Mathematical Biosciences*, 98(2):185–200, 1990.

[20] J. Hein. A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution*, 36(4):396–405, 1993.

[21] J. Hein, M. H. Schierup, and C. Wiuf. *Gene Genealogies, Variation and Evolution.* Oxford University Press, 2005.

[22] J. Hey. Human mitochondrial DNA recombination: can it be true? *Trends in Ecology & Evolution*, 15(5):181–182, 2000.

[23] J. Hey and J. Wakeley. A coalescent estimator of the population recombination rate. *Genetics*, 145(3):833–46, 1997.

[24] W. G. Hill and A. Robertson. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 33:54–78, 1968.

[25] R. R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23:183–201, 1983.

[26] R. R. Hudson. Two-locus sampling distributions and their application. *Genetics*, 159(4):1805–17, 2001.

[27] R. R. Hudson and N. L. Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences.

Genetics, 111(1):147–64, 1985.

[28] D. H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23:254–267, 2006.

[29] H. Innan and M. Nordborg. Recombination or mutational hot spots in human mtDNA? *Molecular Biology and Evolution*, 19(7):1122–7, 2002.

[30] I. B. Jakobsen and S. Easteal. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Computer Applications in the Biosciences*, 12(4):291–5, 1996.

[31] F. M. Jiggins. The rate of recombination in Wolbachia bacteria. *Molecular Biology and Evolution*, 19(9):1640–3, 2002.

[32] T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In *Mammalian Protein Metabolism, III*, pages 21–132. Academics Press, 1969.

[33] M. Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893–903, 1969.

[34] J. F. C. Kingman. The coalescent. *Stochastic Processes and Their Applications*, 13:235–248, 1982.

[35] A. M. Kretzer and T. D. Bruns. Use of atp6 in fungal phylogenetics: an example from the boletales. *Molecular Phylogenetics and Evolution*, 13(3):483–92, 1999.

[36] M. K. Kuhner, J. Yamato, and J. Felsenstein. Maximum likelihood estimation of recombination rates from population data. *Genetics*, 156(3):1393–401, 2000.

[37] E. D. Ladoukakis and E. Zouros. Recombination in animal mitochondrial DNA: evidence from published sequences. *Molecular Biology and Evolution*, 18(11):2127–31, 2001.

[38] W. J. Le Quesne. A method of selection of characters in numerical taxonomy. *Systematic Zoology*, 18(2):201–205, 1969.

[39] R. C. Lewontin. The interaction of selection and linkage. i. general considerations; heterotic models. *Genetics*, 49:49–67, 1964.

[40] D. Martin and E. Rybicki. RDP: detection of recombination amongst aligned sequences. *Bioinformatics*, 16(6):562–3, 2000.

[41] Y. Martin, G. Gerlach, C. Schlotterer, and A. Meyer. Molecular phylogeny of european muroid rodents based on complete cytochrome b sequences. *Molecular Phylogenetics and Evolution*, 16(1):37–47, 2000.

[42] F. J. Massey. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

[43] J. Maynard Smith. Analyzing the mosaic structure of genes. *Journal of Molecular Evolution*, 34(2):126–9, 1992.

[44] J. Maynard Smith and N. H. Smith. Recombination in animal mitochondrial DNA. *Molecular Biology and Evolution*, 19(12):2330–2, 2002.

[45] G. McGuire and F. Wright. Topal 2.0: improved detection of mosaic sequences within multiple alignments. *Bioinformatics*, 16(1367-4803):130–4, 2000.

[46] G. A. McVean. What do patterns of genetic variability reveal about mitochondrial recombination? *Heredity*, 87:613–20, 2001.

[47] G. A. McVean. A genealogical interpretation of linkage disequilibrium. *Genetics*, 162(2):987–91, 2002.

[48] G. A. McVean, P. Awadalla, and P. Fearnhead. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, 160(3):1231–41, 2002.

[49] V. N. Minin, K. S. Dorman, F. Fang, and M. A. Suchard. Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics*, 21(1367-4803):3034–42, 2005.

[50] N. Miyashita and C. H. Langley. Molecular and phenotypic variation of the white locus region in *Drosophila melanogaster*. *Genetics*, 120:199–212, 1988.

[51] S. R. Myers and R. C. Griffiths. Bounds on the minimum number of recombination events in a sample history. *Genetics*, 163(1):375–94, 2003.

[52] R. Nielsen. Site-by-site estimation of the rate of substitution and the correlation of rates in mitochondrial DNA. *Systematic Biology*, 46(2):346–53, 1997.

[53] R. Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154(2):931–42, 2000.

[54] D. Penny and M. Hendy. Estimating the reliability of evolutionary trees. *Molecular Biology and Evolution*, 3(5):403–17, 1986.

[55] G. Piganeau, M. Gardner, and A. Eyre-Walker. A broad survey of recombination in animal mitochondria. *Molecular Biology and Evolution*, 21(12):2319–25, 2004.

[56] D. Posada. Unveiling the molecular clock in the presence of recombination. *Molecular Biology and Evolution*, 18(10):1976–8, 2001.

[57] D. Posada. Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Molecular Biology and Evolution*, 19(5):708–17, 2002.

[58] D. Posada and K. A. Crandall. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 98(24):13757–62, 2001.

[59] D. Posada and K. A. Crandall. The effect of recombination on the accuracy of phylogeny estimation. *Journal of Molecular Evolution*, 54(3):396–402, 2002.

[60] J. Rohayem, J. Munch, and A. Rethwilm. Evidence of recombination in the norovirus capsid gene. *Journal of Virology*, 79(8):4977–90, 2005.

[61] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–25, 1987.

[62] S. Sawyer. Statistical tests for detecting gene conversion. *Molecular Biology and Evolution*, 6(5):526–38, 1989.

[63] S. W. Schaeffer and E. L. Miller. Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. *Genetics*, 135:541–52, 1993.

[64] M. H. Schierup and J. Hein. Consequences of recombination on traditional phylogenetic analysis. *Genetics*, 156(2):879–91, 2000.

[65] M. H. Schierup and J. Hein. Recombination and the molecular clock. *Molecular Biology Evolution*, 17(10):1578–9, 2000.

[66] M. Slatkin. Linkage disequilibrium in growing and stable populations. *Genetics*, 137:331–336, 1994.

[67] M. Slatkin and R. R. Hudson. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, 129(2):555–62, 1991.

[68] P. H. A. Sneath, M. J. Sackin, and R. P. Ambler. Detecting evolutionary incompatibilities from protein sequences. *Systematic Zoology*, 24(3):311–332, 1975.

[69] Y. S. Song and J. Hein. On the minimum number of recombination events in the evolutionary history of DNA sequences. *Journal of*

Mathematical Biology, 48(2):160–186, 2004.

[70] S. Suerbaum, J. Maynard Smith, K. Bapumia, G. Morelli, N. H. Smith, E. Kunstmann, I. Dyrek, and M. Achtman. Free recombination within *Helicobacter pylori*. *Proceedings of the National Academy of Sciences of the United States of America*, 95(21):12619–24, 1998.

[71] M. Sumida, M. Ogata, and M. Nishioka. Molecular phylogenetic relationships of pond frogs distributed in the palearctic region inferred from DNA sequences of mitochondrial 12s ribosomal RNA and cytochrome b genes. *Molecular Phylogenetics and Evolution*, 16(2):278–85, 2000.

[72] D. L. Swofford. *PAUP*. Phylogenetic Analysis using Parsimony (*and Other Methods)*. Sinauer Associates, Sunderland, Massachusetts, 1998.

[73] F. Tajima. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123(3):585–95, 1989.

[74] A. D. Tsaousis, D. P. Martin, E. D. Ladoukakis, D. Posada, and E. Zouros. Widespread recombination in published animal mtDNA sequences. *Molecular Biology and Evolution*, 22(4):925–33, 2005.

[75] T. Uzzell and K. W. Corbin. Fitting discrete probability distributions to evolutionary events. *Science*, 172:1089–1096, 1971.

[76] J. D. Wall. A comparison of estimators of the population recombination rate. *Molecular Biology and Evolution*, 17(1):156–63, 2000.

[77] G. F. Weiller. Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. *Molecular Biology and Evolution*, 15(3):326–35, 1998.

[78] B. S. Weir and W. G. Hill. Nonuniform recombination within the human beta-globin gene cluster. *American Journal of Human Genetics*, 38(5):776–781, 1986.

[79] C. Wiuf, T. Christensen, and J. Hein. A simulation study of the reliability of recombination detection methods. *Molecular Biology and Evolution*, 18(10):1929–39, 2001.

[80] C. Wiuf and J. Hein. The coalescent with gene conversion. *Genetics*, 155(1):451–62, 2000.

[81] Z. Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, 10(6):1396–401, 1993.

[82] Z. Yang. A space-time process model for the evolution of DNA sequences. *Genetics*, 139(2):993–1005, 1995.

[83] Z. Yang. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences*, 13(5):555–6, 1997.

4.8 Expectation and Variance of Φ_w

The normal approximation to the permutation test requires calculation of the expectation and variance of the Φ_w statistic under permutations of the alignment. This section contains derivations for both the mean and variance and outlines how to compute both values efficiently. Again, assume that the proportion of informative sites is q and let w be a fixed width (in bases). Throughout this section, let $k = wq$.

Let $M = (M_{i,j})$ be a given $n \times n$ refined incompatibility matrix. Note that M is symmetric. Let $I = \{1, ..., n\}$ be an index set. Let σ be any permutation of the index set, and define a permutation of the matrix as $\sigma(M) = (M_{\sigma(i),\sigma(j)})$.

Define the sample space Ω by $\Omega = \{\sigma(M) : \sigma \in S_n\}$. Assume that every permutation σ is equally likely. Define an $n \times n$ random matrix $X : \Omega \to \mathbb{R}^{n \times n}$ by $X = \sigma(M)$. Note that X is symmetric, a fact that is used throughout these pages without further mention.

Define for all $1 \leq i \leq n$: $f_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} M_{i,j}$ and $g_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} M_{i,j}^2$.

Also define $u = \sum_{i=1}^{n} f_i$, $v = \sum_{i=1}^{n} g_i$, and $w = \sum_{i=1}^{n} (f_i)^2$

Lemma 4.1. *Let X be a random matrix. Then for any arbitrary but distinct* $\{i, j, k, l\}$

$$\begin{aligned}
\mathrm{E}[X_{i,j}] &= \frac{(n-2)!}{n!}u \\
\mathrm{E}[X_{i,j}^2] &= \frac{(n-2)!}{n!}v \\
\mathrm{E}[X_{i,j}X_{i,k}] &= \frac{(n-3)!}{n!}(w-v) \\
\mathrm{E}[X_{i,j}X_{k,l}] &= \frac{(n-4)!}{n!}(u^2+2v-4w)
\end{aligned}$$

Proof. Note that a permutation σ of I can be viewed as mapping to $I \rightarrow I$. Denote the value of $\sigma(i)$ by σ_i. The total number of permutations is then $n!$. The number of permutations that have m distinct elements fixed in some mapping is $(n-m)!$ (e.g. $\sigma(a_1) = b_1, \sigma(a_2) = b_2, ..., \sigma(a_m) = b_m$). Since every permutation is equally likely the probability of such a permutation is

$$\frac{(n-m)!}{n!}$$

Note that every distinct pair (i, j), $i \neq j$ can be mapped to any distinct pair (a, b), $a \neq b$, by some σ. Note also that $\Pr[X_{i,j} = M_{a,b}] = \Pr[\sigma_a = i \wedge \sigma_b = j]$. Finally, for notational convenience the summation $\sum_{a=1}^{n}$ is written as $\sum_{a}$.

Hence:

$$\begin{aligned}
\mathrm{E}[X_{i,j}] &= \sum_a \sum_{b \neq a} M_{a,b} \Pr[\sigma_a = i \wedge \sigma_b = j] \\
&= \sum_a \sum_{b \neq a} M_{a,b} \frac{(n-2)!}{n!} \\
&= \frac{(n-2)!}{n!} u \\
\mathrm{E}[X_{i,j}^2] &= \sum_a \sum_{b \neq a} M_{a,b}^2 \Pr[\sigma_a = i \wedge \sigma_b = j] \\
&= \frac{(n-2)!}{n!} v \\
\mathrm{E}[X_{i,j} X_{i,k}] &= \sum_a \sum_{b \neq a} \sum_{c \neq a,b} M_{a,b} M_{a,c} \Pr[\sigma_a = i \wedge \sigma_b = j \wedge \sigma_c = k] \\
&= \frac{(n-3)!}{n!} \sum_a ((f_a)^2 - g_a) \\
&= \frac{(n-3)!}{n!} (w - v) \\
\mathrm{E}[X_{i,j} X_{k,l}] &= \sum_{a=1} \sum_{b \neq a} \sum_{c \neq a,b} \sum_{d \neq a,b,c} M_{a,b} M_{c,d} \Pr[\sigma_a = i \wedge \sigma_b = j \wedge \sigma_c = k \wedge \sigma_d = l)] \\
&= \frac{(n-4)!}{n!} \left((\sum_a f_a)^2 + \sum_a (2g_a - 4(f_a)^2) \right) \\
&= \frac{(n-4)!}{n!} (u^2 + 2v - 4w)
\end{aligned}$$

□

Consider the statistic Φ_w defined on a random matrix X as:

$$\Phi_w = \frac{2}{k(2n-k-1)} \sum_{j=1}^{k} \sum_{i=1}^{n-j} X_{i,i+j}.$$

Define (for $1 \leq a, b \leq n$):

$$P_k = \{(a,b) : a < b \leq a + k\}.$$

Note that

$$|P_k| = (n-1) + (n-2) + ... + (n-k) = \frac{k(2n-k-1)}{2}.$$

Then:

$$\Phi_w = \frac{1}{|P_k|} \sum_{(a,b) \in P_k} X_{a,b}.$$

Theorem 4.1. *The expectation and variance of* Φ_w *can be written as (for* $n \geq 2k$*):*

$$\mathrm{E}[\Phi_w] = \frac{(n-2)!}{n!}(u)$$
$$\mathrm{Var}[\Phi_w] = c_1 u^2 + c_2 v + c_3 w$$

where

$$c_1 = \frac{2}{3} \frac{27kn - 18k^2 + 28k^2n - 21kn^2 - 9k + 5n - 9k^3 - 11n^2 + 6n^3 + 6k^3n - 4k^2n^2}{k(k+1-2n)^2(n-1)^2(n-2)(n-3)n^2}$$
$$c_2 = \frac{2}{3} \frac{39kn - 14k^2 + 8k^2n - 15kn^2 - 21k + 19n + 3k^3 - 21n^2 + 6n^3 - 4}{k(k+1-2n)^2n(n-1)(n-2)(n-3)}$$
$$c_3 = -\frac{4}{3} \frac{-18kn - 2k^2n + 16k^2 + 6n^2 - 10n + 2 + 15k + 3k^3}{k(k+1-2n)^2n(n-1)(n-2)(n-3)}$$

Moreover, both $\mathrm{E}[\Phi_w]$ *and* $\mathrm{Var}[\Phi_w]$ *can be calculated in* $O(n^2)$ *time.*

Proof. The expectation is straightforward:

$$\mathrm{E}[\Phi_w] = \frac{1}{|P_k|} \sum_{(a,b)\in P_k} \mathrm{E}[X_{a,b}] = \frac{(n-2)!}{n!} u$$

The variance is a little more involved:

$$\begin{aligned}\mathrm{Var}[\Phi_w] &= \mathrm{Var}\left[\frac{1}{|P_k|} \sum_{(a,b)\in P_k} X_{a,b}\right] \\ &= \frac{1}{|P_k|^2}\left(\sum_{(a,b)\in P_k} \mathrm{Var}[X_{a,b}] + 2 \sum_{((a,b),(c,d))\in Q_k} \mathrm{Cov}[X_{a,b}X_{c,d}]\right)\end{aligned}$$

where

$$Q_k = \{((a,b),(c,d)) \in P_k \times P_k : (a,b) \prec (c,d)\}$$

and $\prec$ denotes standard lexicographical ordering.

Note that Q_k can be partitioned into 2 disjoint sets $Q_{k,0}$ and $Q_{k,1}$ where $Q_{k,m} = \{((a,b),(c,d)) \in Q_k : |\{a,b\} \cap \{c,d\}| = m\}$ $\Big($by definition Q_k does not contain pairs of the type $((a,b),(a,b))\Big)$. One way to determine $Q_{k,1}$ is to set up a recurrence.

Note that:

$$P_1 = \{(1,2),(2,3)...(n-1,n)\}$$

so that

$$Q_{1,1} = \{((a,a+1),(a+1,a+2)) : 1 \le a \le n-2\}$$

Hence $|Q_{1,1}| = (n-2)$.

Next let $((a_1,a_2),(a_3,a_4)) \in Q_k - Q_{k-1}$. Then at least one $(a_1,a_2) = (a,a+k)$ or $(a_3,a_4) = (a,a+k)$ must be true. Consider the four subcases:

Case 1: $((a,b),(a,a+k))$ where $1 \leq a \leq n-k$ and $a < b < a+k$. There are precisely $(n-k)(k-1)$ terms of this type.

Case 2: $((a,a+k),(b,a+k))$ where $1 \leq a \leq n-k$ and $a < b < a+k$. Again, there are precisely $(n-k)(k-1)$ terms of this type.

Case 3: $((a,a+k),(a+k,b))$ where $1 \leq a \leq n-k$ and $a+k < b \leq \min(a+2k,n)$. For $n \geq 2k$ there are $(k)((n-k)-k)+(k)(k-1)/2$ such terms.

Case 4: $((b,a),(a,a+k))$ where $1 \leq a \leq n-k$ and $\max(1,a-k) \leq b < a$. For $n \geq 2k$ there are again $(k)((n-k)-k)+(k)(k-1)/2$ such terms.

Cases 3 and 4 can coincide for $n \geq 2k$ when $|a-b| = k$. All other combination of cases are disjoint. There are precisely $(n-k)-k$ such coincidences. This gives the following recurrence for $Q_{k,1}$:

$$Q_{k,1} = 2(n-k)(k-1)+(k-1)(k)+(2k-1)(n-2k)+Q_{k-1,1}$$

$$Q_{1,1} = n-2$$

The recurrence can be solved by standard techniques resulting in:

$$Q_{k,1} = 2k^2n - \frac{5}{3}k^3 - kn + \frac{2}{3}k - k^2$$

Note that $|Q_k| = \binom{|P_k|}{2}$. Since Q_k is the disjoint union of $Q_{k,0}$ and $Q_{k,1}$, then

$$|Q_{k,0}| = |Q_k| - |Q_{k,1}|$$

The variance of Φ_w can then be written as

$$\begin{aligned}\mathrm{Var}[\Phi_w] &= \frac{1}{|P_k|^2}\Bigg(\sum_{(a,b)\in P_k}\mathrm{Var}[X_{a,b}] + 2\sum_{((a,b),(c,d))\in Q_{k,0}}\mathrm{Cov}[X_{a,b}X_{c,d}] + \\ &\quad 2\sum_{((a,b),(c,d))\in Q_{k,1}}\mathrm{Cov}[X_{a,b}X_{c,d}]\Bigg) \\ &= \frac{1}{|P_k|^2}\Big(|P_k|\mathrm{Var}[X_{a,b}] + 2|Q_{k,0}|\mathrm{Cov}[X_{a,b}X_{c,d}] + 2|Q_{k,1}|\mathrm{Cov}[X_{a,b}X_{a,c}]\Big)\end{aligned}$$

Noting that $\mathrm{Cov}[X_{a,b}X_{c,d}] = \mathrm{E}[X_{a,b}X_{c,d}] - \mathrm{E}[X_{a,b}]\mathrm{E}[X_{c,d}]$ and $\mathrm{Var}[X_{a,b}] = \mathrm{E}[X_{a,b}^2] - \mathrm{E}[X_{a,b}]^2$, the constants c_1, c_2 and c_3 can be solved for using the relations from the previous Lemma. Since the quantities u, v and w can be computed in $O(n^2)$ time, so can the variance and expectation. □

4.9 Additional parameters for coalescent simulation

The rate of recombination is here referred to as $\rho = 4Nrt$ where r is the per-base recombination rate and t is the sequence length. Here N was set to 1000 (diploid population), t was set to 1000 as well, and r solved for accordingly.

For population growth $\rho^\dagger$ was obtained so that the expected number of recombinations was equal under scenarios (i.e. $\mathrm{E}_{\beta=5000}[R(m)] = \mathrm{E}_{\beta=0}[R(m)]$) where $R(m)$ is the number of recombinations for a sample of size m [79] and $\beta = Nb$ where b is the population growth rate per generation [79]. The expected number of recombinations for $\beta = 0$ can be found by the following formula [27]:

TABLE 4–7: Conversion of the rate of recombination ρ between $\beta = 0$ and $\beta = 5000$

Sample Size	E$[R(m)]$	ρ	
		$\beta = 0$	$\beta = 5000$
$m = 5$	2.08	1	550
$m = 10$	2.83	1	400
$m = 15$	3.25	1	325
$m = 25$	3.78	1	250
$m = 50$	4.48	1	175

$$\mathrm{E}_{\beta=0}[R(m)] = \rho \sum_{j=1}^{m-1} \frac{1}{j}$$

Table $4 - -7$ shows the values used for $\rho = 1$ (when $\beta = 0$). For values of $\rho > 1$, (e.g. $\rho = 2$) one can simply double the values in the table.

Similarly, the rate of mutation is here referred to as $\theta = 4N\mu t$ where μ is the per-base mutation rate and t is the sequence length. Under a Jukes-Cantor model if $\beta = 0$ then

$$\theta = t\frac{3p}{3 - 4p}$$

[79]. This allows θ to be found for a fixed amount of sequence diversity p. For $\beta = 5000$ the appropriate value of θ was found by simulation. The values used are shown in Table 4–8.

TABLE 4–8: Conversion of the rate of mutation θ between $\beta = 0$ and $\beta = 5000$

Diversity	θ	
	$\beta = 0$	$\beta = 5000$
$p = 1\%$	10.1	6600
$p = 5\%$	53.6	33000
$p = 10\%$	115.4	68000
$p = 15\%$	187.5	106000
$p = 25\%$	375	193600

CHAPTER 5

Recombination shapes the evolution of FIV in a wild population of cougars

5.1 Background

This chapter applies the statistical test developed in Chapter 4 to understand recombination within a genome level data set. Recombination is shown to play an important role in the evolution of FIV in the wild.

5.2 Abstract

Recombination confounds phylogenetic analysis and plays an important role in viral evolution. But ascertaining where recombination has occurred and which sequences may be recombinant is a difficult task. Here we use a recently developed statistical test to perform exploratory analysis of recombination in 14 feline immunodeficiency virus (FIV) complete genomes taken from a wild population of cougars. The novelty of the approach is that different genomic regions are tested individually for recombination (with multiple test correction) to identify recombinant regions, before using traditional approaches such as similarity plots. We identify three strains derived from recombinant events and phylogenetic incongruence tests

confirm their mosaic nature. Previous studies of FIV from wild cougars have focused on the epidemiology of the virus and the population structure. The results suggest that recombination has played an important role in the evolution of FIV for this wild population of cougars.

5.3 Introduction

Recombination creates new genotypes by combining genetic material from distinct lineages. A major force in viral evolution, recombination can increase viral pathogenicity and is a principal component in creating genetic diversity [2, 24]. Indeed, the astounding variety of viral forms and their central role in evolution is in part a result of the extensive recombination in the viral world [12]. From a practical standpoint, recombination among viral strains may complicate vaccine development, for instance [24].

Broadly speaking, recombination complicates phylogenetic analysis [30]. If recombination has occurred the time to the most recent common ancestor to be biased when performing phylogenetic analysis [30]. Moreover, recombination creates signals consistent with exponential growth [30].

Due to the impact of recombination on phylogenetic analysis, an important question in viral evolutionary analysis is to detect whether recombination has occurred [2, 24]. Because of the confounding effects of substitution rate heterogeneity, rigorous statistical analysis must be used to separate recombination from other processes [1]. Towards this end, a number of approaches have been developed to detect and analyze recombinant sequences [21, 23]. Although there are specifically developed population

genetic approaches for inferring the rate and presence of recombination, these are usually only appropriate when all the strains originate from a single population (in an infected host for instance) [8]. A number of approaches based on phylogenetic principles and summary statistics that are applicable to analysis of divergent strains have also been developed. Despite the fact that phylogenetic approaches such as bootscanning [29] are popular for recombinant analysis, compared to other methods they have low sensitivity and can produce an excess false positives [21, 23]. Moreover recently it has been shown that summary statistics such as Max Chi Squared [17], although hitherto considered the most powerful methods [21] may produce an excess of false positives in certain cases and fail to detect recombination in other circumstances [8].

A new statistical approach called the Phi statistic has been recently developed to analyze recombinant sequences. Given a sequence alignment, the Phi test accurately determines whether recombination has occurred or not. The novelty of the method is that it is very sensitive, does not produce false positives in excess and can be applied to divergent or closely related strains. These characteristics make it ideal to distinguish between recurrent mutation and recombination in viral genomes [8].

However, the Phi test as proposed in previous work simply determines whether or not recombination has occurred within a set of sequences [8]. The original Phi test can thus be thought of as a global test for recombination, and should be used as a first step in analyzing sequences. But a more

fine-scale approach is needed in order to determine where recombination may have occurred. Here we propose using the Phi statistic to test smaller regions for recombination resulting in local tests for recombination. There is a tradeoff however for the gain in knowledge of where recombination may occur. Firstly, testing smaller regions for recombination results in less sensitivity, since recombination is easier to detect in longer sequences [34]. Secondly, multiple test correction must be applied since instead of a single global test for recombination, a number of tests on smaller regions are being applied making the procedure much more conservative. This could potentially result in a situation where there is global evidence for recombination, but it is difficult to pinpoint the exact location of the recombinant signal.

Testing local regions for recombination accomplishes two goals however. Firstly, areas that contain a recombinant signal (in other words potential breakpoints) can be easily identified. Secondly, phylogenetic trees can be built between the areas of recombinant origin allowing putative recombinant sequences to be identified. This allows hypothesis generation about which sequences may or may not be recombinant. Additional data analysis can then be done efficiently using techniques such as similarity plots [16]. Here we apply the Phi test and other approaches to study recombination in FIV, a well-known pathogen.

Feline immunodeficiency virus (FIV) is a lentivirus of the family Retrovidae and is similar to human immuno-deficiency virus (HIV) [20]. Indeed, the similarity between FIV in domestic cats and HIV has prompted

suggestions of using FIV as a model for AIDS studies as well as vaccine development [4, 15]. Although FIV was first described in domestic cats [20], it has been found in a wide variety of wild felines including lions (Panthera leo) as well as wild cougars (Puma concolor) [7, 18].

The clinical symptoms of domestic cats infected with FIV contrast strongly to the symptoms exhibited by wild feline species infected with FIV, however. In domestic cats, FIV infection can lead to immune dysfunction as well as behavioral problems [6, 9]. On the other hand FIV in wild felines does not appear to lead to disease, possibly due to a lengthy co-adaptation of virus and host [6, 9].

Recombination across genetically distinct strands is known to play a strong role in the evolution of HIV [26]. Recombination between divergent strains ultimately depends on the presence of co-infected hosts [26]. The exact mechanisms of recombination in HIV are not fully understood, but recombination arises in a co-infected cell during reverse transcription [11]. Recombination across divergent strains has also been observed across broader classes of viruses, such as the Primate Lentivirus (PLV) strains which include HIV-1 and HIV-2 as subgroups [28].

Recombination in FIV has not been studied to the same extent as recombination in HIV or PLV. Previous work on recombination in FIV has generally focused on recombination of the virus in domestic cats. For instance, an analysis of a worldwide population of FIV infected domestic

cats revealed the presence of multiple recombinant subtypes [3]. A geographically focused study of infected domestic cats in Ontario, Canada suggested the presence of a circulating recombinant form enzootic to that region [25]. In wild cougar FIV however, previous studies have not firmly established the presence of recombination across genetically distinct strains [6]. A typical approach for detecting recombination in these studies is to rely upon incongruent gene trees for the sequenced strains, typically determined by incongruence using *gag*, *env* or *pol* genes. In this study, we focus on analyzing entire viral genomes of FIV strains, to determine the extent of recombination within the strains.

In this paper, we use both the global and local Phi statistical approach as an overall guide to predict where recombination may have occurred. For the FIV data set considered here, the Phi statistic combined with exploratory phylogenetic analysis identifies three potential recombinant lineages simultaneously. Further analysis, including similarity plots and phylogenetic incongruence tests, confirm that these major lineages are descendant of a recombinant event. Interestingly, some of the regions of mosaic origin appear to have been involved in more than one recombination event. Since some of the mosaic strains are currently geographically disparate, the results indicate that the geographical dispersal of the observed strains was preceded by recombination events.

5.4 Methods

5.4.1 Data Set

An alignment of fourteen FIV genomes sequences; GC34, MC100, MC121, MC350, YM29A/YF16, JM01/YM137, JF6, YF125, SR631/SR631B, CoLV and PLV (slashes indicate pairs of strains with high sequence similarity) was done by M. Poss (personal communication). The PLV strain is an out-group strain collected from Vancouver island in 1995. YM29A/YF16 were taken from cougars in the Yellowstone area in 1991. There are two other Yellowstone strains are YM137 and YF125. ColV was taken from a cougar in the interior of British Columbia in the mid 1980s. The rest of the strains were collected from various points in the Northwestern United States and Canada. The SR631 and SR631B were collected from a point in SE Wyoming over 1000 kilometers away from the rest of the sequences separated by inhospitable territory.

5.4.2 Exploratory Recombinant Analysis with Phi statistic

Determining the major recombinant events that have shaped all 14 FIV genomes proceeded in a number of major steps. Firstly, the entire alignment was tested for recombination using the Phi statistic [8]. Next, regions containing a recombinant signal (at least one breakpoint) were identified as follows. Local, overlapping regions of 500 base pairs (adjacent regions differed by 25 base pairs) were tested for recombination using the Phi statistic, using the Profile program [8]. To ensure significance in the

presence of multiple tests, Hommel's [14] modified Bonferroni procedure was applied to the p-values.

Exploratory neighbor-joining trees were built for each of the regions that did not contain a recombinant signal using PAUP* to identify recombinant lineages [27, 32]. A larger, reference maximum likelihood tree for the initial non-recombinant part of the alignment was built for the entire set of sequences, based on a best fitting substitution model found with ModelTest [22] and inferred using PAUP* [10, 32] (a heuristic search was performed using tree-bisections and reconnections). All trees were drawn using the program TREEVIEW [19].

5.4.3 Fine-scale recombinant analysis

Given the putative recombinants identified diversity plots were then created between the putative recombinant and its parents using the program Simplot [16]. Simplot was run using a window-size of 200 base pairs and a step size of 20 base pairs. Breakpoints were identified using the putative recombinant and both parents under a maximum likelihood framework using the program LARD [13] and the location of the breakpoints were compared to the regions identified by the Phi statistic.

Finally, phylogenetic trees for the individual partitions (regions that did not contain a breakpoint) were built using the major lineages. The partitions were constructed to compare a putative placement of a recombinant lineage and thus only contained a subset of the taxa. A best fitting nucleotide substitution model for each partition was found using the Akaike

Information Criterion (AIC), implemented in ModelTest [22]. A check was then used to determine whether each partition was free of recombination using the Phi test [8]. Next, based on the inferred best model of substitution, maximum likelihood trees were found using PAUP* [10, 32] (a heuristic search was performed using Tree-Bisections and Reconnections). Finally, for each region, the different possible trees were compared to each other using a one sided Shimodaira-Hasegawa test [31] based on resampling-estimated log likelihoods (RELL) bootstrapping 10000 replicates available in PAUP* [31]. Trees were drawn using the program TREEVIEW [19].

5.5 Results

5.5.1 Identification of recombinant (breakpoint) regions using Phi statistic

Initially all positions of all 14 viral genomes were tested for recombination using the Phi statistic [8]. Overwhelming evidence of recombination within the entire alignment was found, that is $p < 10^{-20}$. Individual regions of 500 base pairs were then tested for recombination using the Phi statistic. This clearly demarcated a number of different recombinant regions (Figure 5–1). Figure 5–1 shows that there are at least six distinct smaller regions exhibiting statistically significant evidence of recombination (regions containing breakpoints) within the 14 FIV genomes.

The strongest local evidence of recombination occurs in the second half of the genomes after base 4500 (Figure 5–1). Based on this information, a phylogenetic tree was constructed based on the first half of the alignment

(sites 1 – 4500) (Figure 5–2). This tree represents an estimate of the actual phylogeny of the FIV sequences. This tree can be used as a reference tree, where sequences in putative recombinant regions that are misplaced with respect to this topology are identified as tentatively recombinant. The tree places JM01/YM137, JF6 and YF125 as a monophyletic clade. The lineage SR631/SR631B is placed a strong out-group to this monophyletic clade.

5.5.2 Exploratory analysis of recombinant regions identified by Phi statistic

To identify which sequences are recombinant seven different phylogenetic trees were built using neighbor-joining [27] based on regions in-between the six statistically significant recombinant regions. These trees are shown in Figure 5–3. Only major lineages are shown. Thus, sequences such as SR631/SR631b are shown by SR631 for simplicity. Note that YM29A and CoLV serve as 'anchor' sequences on one branch, YF125 and PLV serve as 'anchor' sequences on the other branch. The putative recombinants change position relative to these sequences

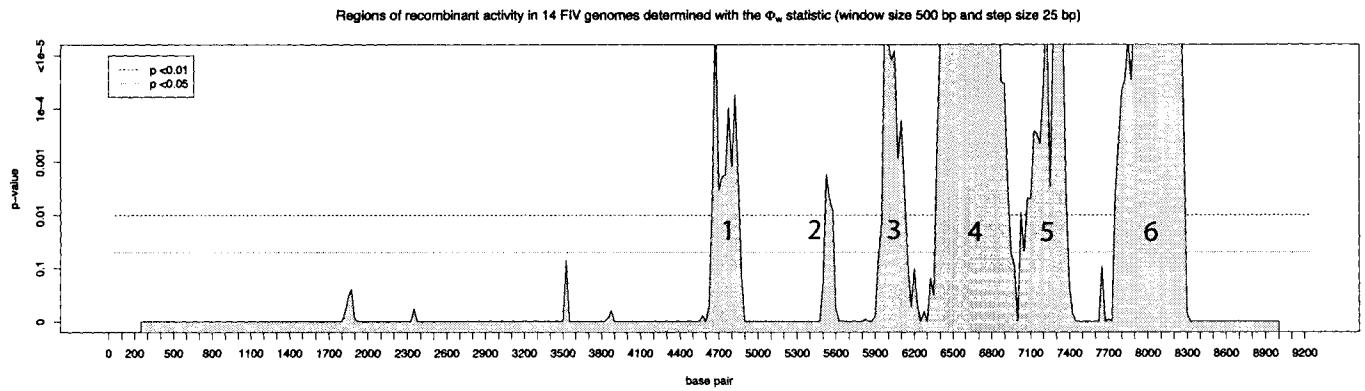


FIGURE 5–1: Statistically significant regions (peaks that contain at least one breakpoint) containing a recombinant signal of all 14 FIV genomes determined by Phi statistic (p values are multiple test corrected using Hommel's modified Bonferroni procedure [14])

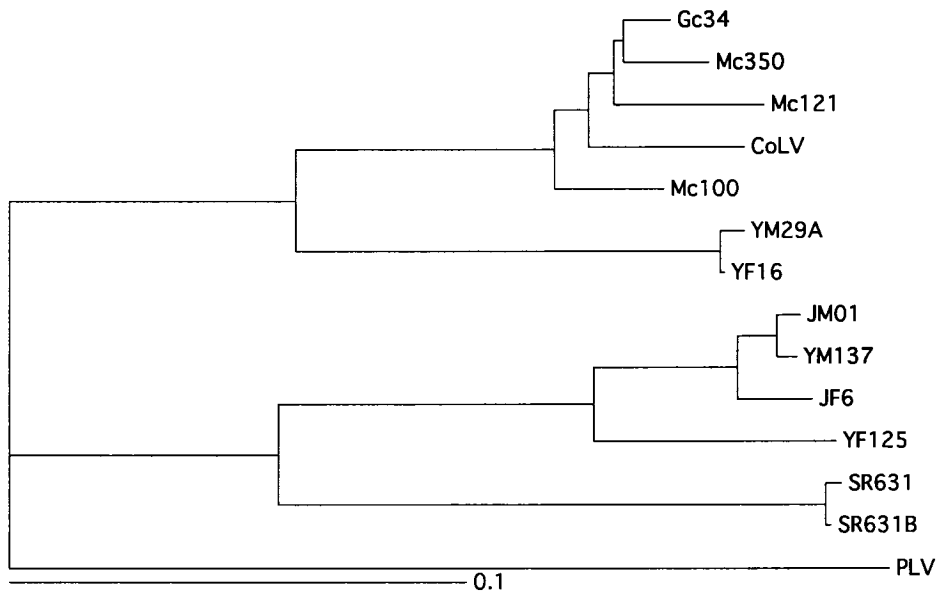


FIGURE 5–2: Maximum likelihood tree inferred using all sequences up to base 4500 (before the first recombinant region or 'peak' in Figure 5–1). A substitution model of GTR+Γ was selected using ModelTest [22] (with Γ equal to 0.26) and was used to infer the tree.

Both the first and last tree display a history consistent with the reference tree (Figures 5–3a and 5–3g compared to Figure 5–2). Note for instance that the second tree (Figure 5–3b) groups JM01 with SR631, in contrast to the first tree that groups JM01 with JF6 and YF125. This placement of JM01 and SR631 together in the second tree indicates that either JM01 or SR631 derives from a recombinant event. But using the 'anchor sequences YF125 and PLV, the second tree suggests that JM01 is derived from a recombinant event since it changes position with respect to both of these sequences.

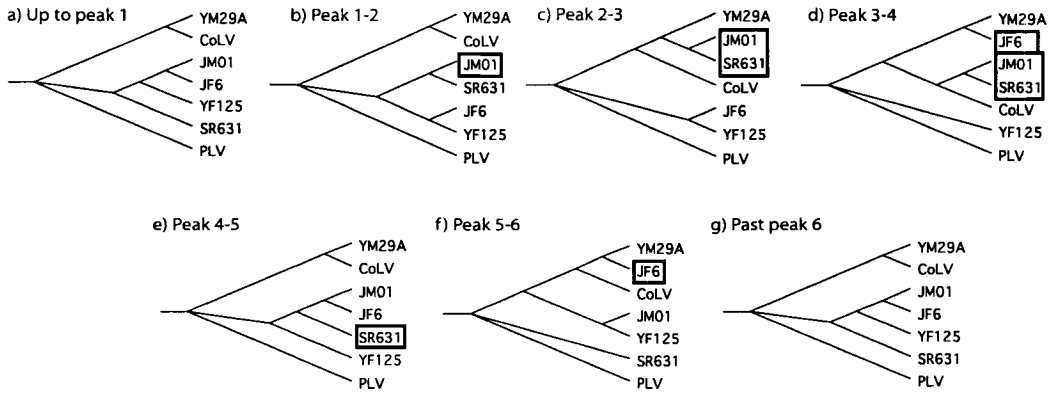


FIGURE 5–3: Exploratory neighbor-joining trees corresponding to the regions in between 'peaks' of Figure 5–1 (only major lineages are shown for readability). Since some peaks consist of many bases, the midpoint of each peak was chosen to delineate the region. Lineages that have boxes around them, namely JM01, SR631 and JF6 appear to be recombinant.

Comparison of the second tree to the third tree (Figure 5–3b and 5–3c), provides preliminary evidence that JM01/SR631 is itself an offshoot of an older recombinant event with the ancestor to YM29A. This suggests that SR631 is also a candidate recombinant sequence (JM01 is already on the list of recombinant sequences). The fourth tree (Figure 5–3d) suggests additionally that JF6 is also recombinant sequence. The fifth and sixth trees (Figure 5–3e and 5–3f) also confirm that all three sequences (JM01, SR631 and JF6) are possible recombinant sequences. However, the exploratory analysis does not precisely identify where the recombination has occurred, partly because the identified recombinant regions (regions containing a breakpoint) are quite large, and there is the possibility of overlapping recombinant events. Nonetheless, the exploratory analysis suggests that

there are at least three sequences that demand further study: JM01, SR631 and JF6.

5.5.3 Fine-scale recombinant analysis and exact breakpoint identification

The exploratory analysis suggested that JF6, SR631 and JM01 were derived from recombinant events. Similarity plots [16] were then constructed using these sequences (Figure 5–4). Similarity plots can be used to refine the hypothesis of recombinant origin for a particular sequence by suggesting where recombination may have occurred. However, similarity plots alone do not provide evidence that this is a recombinant event [1], they merely help refine the hypothesis of which area of the sequence may be recombinant. The information gathered from these plots can be used to test the hypothesis of recombinant origin for particular sequences, using for instance phylogenetic trees (discussed in following section). The details of which sequences were selected for each of the similarity plots along with the major results obtained from the plots are discussed next.

JF6 is a putative recombinant sequence with the ancestral sequence to YM29A as one parent (Figure 5–3d and 5–3f). Unfortunately, the other parent, the ancestral sequence to JM01 (Figure 5–3a and 5–3g for instance) appears to be itself a recombinant sequence (Figure 5–3b). Instead of JM01, an 'anchor lineage, YF125 (Figure 5–3a, Figure 5–2) can be chosen as a non-recombinant lineage similar to JF6. Figure 5–4a shows a similarity plot of JF6, YM29A and YF125. Figure 5–4a suggests that JF6 has a mosaic origin, with two distinct regions of the sequence appearing to originate

from an ancestral sequence closely related to YM29A. The other regions support JF6 closely grouped with YF125. A maximum likelihood procedure, with the program LARD [13] was used to identify the exact breakpoints (Figure 5–4a dashed lines). Note that the breakpoints in Figure 5–4a, fall into peaks of recombinant activity 3, 5 and 6 identified by the Phi statistic (Figure 5–1). It is now possible to use the mosaic regions defined here to test specifically whether there are two distinct phylogenetic trees based on these regions that describe the history of the JF6.

JM01 was also identified as a recombinant sequence with one parent as a sequence closely related to SR631 (Figure 5–3b-d). The 'anchor sequence YF125 was also chosen for similarity plot comparison (Figure 5–3a and 5–3g, Figure 5–2). The similarity plot (Figure 5–4b) suggests one large region where JM01 is much more similar to SR631 rather than YF125. This suggests that potentially one region of JM01 is much more closely related to SR631 than YF125, contradicting the reference tree in Figure 5–4. The maximum likelihood breakpoints found using LARD [13] are shown with dashed lines. Again, the breakpoints shown in Figure 5–4b, correspond to regions of recombinant activity identified by the Phi statistic (Figure 5–1, peaks 1 and 4).

The trees in Figures 5–3c and 5–3d suggest that the mosaic history of JM01 with respect to YM29A deserves further exploration. The similarity plot of JM01, YM29A and YF125 is given in Figure 5–4d. This Figure suggests a region that closely groups JM01 and YM29A, again contradicting

the reference tree in Figure 5–2. Interestingly, this region is fully contained within the JM01 and SR631 putative recombinant region (Figure 5–4b and 5–4d). The breakpoints as identified with LARD [13] are shown with dashed lines and correspond to peaks 2 and 4 identified with the Phi statistic (Figure 5–1).

Finally, exploratory analysis indicated that SR631 may itself be a recombinant sequence that is related to YM29A (Figure 5–3c and 5–3d). However, in the reference tree SR631 appears to be closely related to YF125 (Figure 5–3a, 5–3g and 5–2). The similarity plot (Figure 5–4c) shows that indeed there is a region in SR631 that has higher sequence similarity to YM29A than to YF125. Interestingly, this region corresponds to the region in JM01 that is similar to YM29A. This indicates that SR631 and JM01 have a shared history for this recombination event.

5.5.4 Data partitions

The previous analyses provide a number of potential recombinant lineages JF6, JM01 and SR631. Furthermore, exact hypothetical locations of recombination events were identified. However, similarity plots do not conclusively demonstrate incongruent phylogenetic histories [1]. In order to firmly demonstrate that JF6, JM01 and SR631 are recombinant sequences, topological incongruence tests must be performed showing that different regions of the sequence support different origins of the potentially mosaic sequences.

In this case, there are three potential recombinant lineages JF6, SR631 and JM01. In order to simplify the analysis the hypothetically recombinant sequence JF6 is analyzed separately from SR631 and JM01. Five different partitions of the alignment and the sequences were identified based upon the breakpoint identification and previous analyses (Table 5–1). Note that Table 5–1 also includes both closely related lineages SR631B and SR631. Likewise YM137, which is closely related to JM01 is included in the analysis. These lineages were excluded from earlier analysis in order to simplify the problem. To simplify analysis, two of the partitions (Partition-A and Partition-B) only involve JF6 and do not include sequences from possible recombinants SR631 and JM01. This allows full focus on JF6 rather than other sequences as well. Partition-A consists of all the regions where JF6 appears to have placement consistent with the reference tree, that is the region where JF6 appears to group closely with YF125 (Figure 5–2 and Figure 5–4a). Likewise, Partition-B consists of both regions where JF6 appears to group closely with YM29A (Figure 5–4a).

Partition-C, Partition-D and Partition-E (Table 5–1), exclude the potential recombinant JF6. Partition-C consists of the regions where JM01, SR631 and YF125 have a relationship consistent with the reference tree (Figure 5–2 and Figure 5–4b,c). Partition-D and Partition-E consist of the regions where JM01 and SR631 appear to have shared histories (Figure 5–4b). More precisely, Partition-D consists of the regions where JM01, SR631 and YF125 appear to be closely related, whereas Partition-E consists of the

TABLE 5–1: Description of data partitions including nucleotide locations and substitution models.

Name of partition	Recombinant sequence(s)	Location[a,b]	Recomb-[c] ination?	Substitution model [d]	α[d]	P_{inv}[d]
Partition-A	JF6	$1-5996$, $6273-7237$, $8120-9270$	1.6×10^{-4}*	GTR+Γ+I	2.7	.50
Partition-B	JF6	$5997-6273$, $7238-8119$	2.0×10^{-5}*	GTR+I	-	.52
Partition-C	JM01/YM137 SR631/SR631B	$1-4700$, $6858-9270$	3.0×10^{-7}*	GTR +Γ +I	2.1	.48
Partition-D	JM01/YM137 SR631/SR631B	$4701-5452$, $6462-6857$	0.17	GTR+Γ	.31	-
Partition-E	JM01/YM137 SR631/SR631B	$5453-6461$	1.9×10^{-5}*	GTR+Γ+I	2.9	.46

[a] Breakpoints estimated by maximum likelihood using LARD [13].
[b] Distinct regions are concatenated together within each partition.
[c] p-values for recombination within each partition using Phi test [8].
[d] Estimated using ModelTest based on AIC.
* $p < .05$

regions where JM01, SR631 and YM29A appear to be closely related (Figure 5–4b-d). Each partition was tested for recombination using the Phi statistic. The Phi statistic indicated that each of the partitions except Partition-D still contained a signal for recombination (Table 5–2). This suggests that there is evidence for recombination within each of the partitions. A best fitting nucleotide substitution model was also found for each partition (Table 5–1). In each case, the General Time Reversible (GTR) model was chosen as the best fitting nucleotide substitution model and except in one case a Gamma model [33, 35] for substitution rate heterogeneity.

TABLE 5–2: Phylogenetic incongruence between different partitions

Name of partition / tree[a]	Recombinant sequence(s)	Possible trees[a]	- Log likelihood[b]	p-values SH Test[c]
Partition-A	JF6	Partition-A	27504	-
		Partition-B	28385	0.000*
Partition-B	JF6	Partition-B	3942	-
		Partition-A	3970	0.012*
Partition-C	JM01/YM137,	Partition-C	26915	-
	SR631/SR631B	Partition-D	27062	0.002*
		Partition-E	27935	0.000*
Partition-D	JM01/YM137,	Partition-D	4567	-
	SR631/SR631B	Partition-C	4618	0.007*
		Partition-E	4633	0.001*
Partition-E	JM01/YM137,	Partition-E	4041	-
	SR631/SR631B	Partition-C	4079	0.000*
		Partition-D	4076	0.000*

[a] Maximum likelihood trees for each partition are shown in Figures 5–5 and 5–6.
[b] Evaluated using best nucleotide substitution model for each partition (Table 5–1).
[c] Estimated using 10000 RELL samples.
* $p < .05$

5.5.5 Phylogenetic incongruence confirmation of mosaic sequences

All the previous analyses allow tests to be performed to show that JF6, SR631/SR631B and JM01/YF16 are indeed recombinant. Five different maximum likelihood phylogenetic trees were built using Partition-A to Partition-E, with the substitution model of best fit (Table 5–1). The maximum likelihood trees of Partition-A and Partition-B differ in their

placement of JF6 (Figure 5–5a and 5–5b). A Shimodaira-Hasegawa (SH) test [31] confirms that JF6 has two statistically significant histories based on different regions of the sequence (p-value 0.01, Table 5–2). Partition-A supports the reference tree, where JF6 is closely related to YF125. Partition-B supports JF6 as closely related to YM29A. Note that these are two very distinct histories, demonstrating that JF6 originated from a mosaic sequence.

The maximum likelihood tree topologies supported by Partition-C, Partition-D and Partition-E are given in Figure 5–6a , 5–6b and 5–6c respectively. Partition-C supports the reference topology which places JM01/YM137 with YF125 as a monophyletic clade to the exclusion of SR631/SR631B (Figure 5–2 and Figure 5–6a). Conversely, Partition-D and Partition-E support JM01/YM137 forming a monophyletic clade with SR631/SR631B (Figure 5–6b and 5–6c). Furthermore, Partition-E also supports the monophyletic grouping of YM29A/YF16 with the JM01/YM137/SR631/SR631B monophyletic clade (Figure 5–6c). The SH shows that the tree found for Partition-C provides a significantly better fit to the data in Partition-C than both the alternative topologies found for Partition-D and Partition-E (p-value < 0.01, Table 5–2). Likewise, the SH test shows that the this is also true for the trees in Partition-D and Partition-E (p-value < 0.01, Table 5–2). This demonstrates that JM01/YM137 as well as SR631/SR631B all originated from mosaic sequences.

5.6 Discussion

The local Phi statistical approach was combined with phylogenetic tree building, in order to generate hypotheses about which sequences may be recombinant. From a methodological standpoint, testing local regions for recombination (with multiple test correction) provides a statistically coherent step to ascertain regions that contain breakpoints. Testing for recombination allows better ascertainment of where recombination (that is breakpoints) occur although it does not identify precisely which sequences are recombinant, or the boundaries of recombination. Tentative recombinant sequences can then be identified by also performing exploratory phylogenetic analysis, however.

After tentative recombinant sequences were identified, traditional similarity plots were used to perform further exploration of the data. The precise locations of recombination breakpoints were then estimated using LARD [13]. The breakpoint locations were then used to confirm the phylogenetic discordance of different genomic regions. Unfortunately, subsequent tests for recombination within each of the identified partitions suggested that a recombinant signal was still present within each partition (Table 5–2). This evidence suggests that either the breakpoints were mis-estimated or that other recombinant events were not properly identified. Nonetheless, the ability to test for recombination (using the Phi test) provides an important diagnostic step in this procedure.

Three major strains in the FIV data set JF6, JM01/YM137 and SR631/SR631B were identified as mosaic strains using the Phi statistic [8] as well exploratory phylogenetic analysis. Fine-scale analysis of these strains using similarity plots [16] and phylogenetic incongruence confirmed that these strains descended from recombinant events. The final breakpoints found for the recombinant events corresponded with the recombinant regions tentatively identified by the Phi statistic. The mosaic nature of each of the lineages suggests overlapping recombinant events with a complex interpretation, which merits further discussion.

The first strain, JF6 is derived from a mosaic sequence with most of the sequence originating from a sequence ancestral to YF125 (Figure 5–4a, Figure 5–5a). However, two small portions of the JF6 sequence originate from a sequence ancestral to YM29A/YF16 instead of YF125 (Figure 5–4a, Figure 5–5b). This suggests that the strain ancestral to YM29A/YF16 recombined with a strain ancestral to YF125 to produce a recombinant strain, which led to the sequence JF6.

Another strain, JM01/YM137 also appears to be derived from a recombinant event (Figure 5–4b, Figure 5-6a-c). For the most part, JM01/YM137 appears to generally form a monophyletic clade that contains YF125 but does not contain SR631/SR631B (Figure 5–2, Figure 5–6a). However, a significant portion of the JM01/YM137 strain appears as monophyletic with SR631/SR631B (Figure 5–4b, 5–6b-c) without YF125. This suggests that the JM01/YM137 strain has two different origins. One part of the strain

derives from a sequence ancestral to YF125, whereas another part of the strain derives from a sequence ancestral to SR631/SR631B.

Given that JM01/YM137 ultimately derived from one strain ancestral to SR631/SR631B and another strain ancestral to YF125, a further complication is introduced in Figure 5–6c. In Figure 5–6c, JM01/YM137 and SR631/SR631B form a monophyletic clade with YM29A/YF16 instead of YF125 like in Figure 5–6a-b. Similarity plots (Figure 5–4c-d) show the same phenomenon; that both strains JM01 and SR631 appear at one point to share much more genetic similarity with YM29A rather than YF125. Since the regions of conversion of both JM01 and SR631 are identical, this suggests the recombination events are not independent. The simplest explanation is that the SR631/SR631B lineage derived from a recombinant event, in other words has two distinct origins. One part of the SR631/SR631B strain shares a common ancestor with YF125, whereas another part of the strain shares an ancestor with YM29A/YF16. This is illustrated in Figure 5–6a-c.

Thus Figure 5–6c illustrates the result of two historically recombinant events. The first event created a recombinant ancestor of SR631/SR631B that derived part of its genetic sequence from YM29A/YF16. The strain ancestral to JM01/YM137 at some point then recombined with this mosaic lineage leading to the situation observed in Figure 5–6b-c. This region has undergone multiple recombinant events in the past, suggesting that susceptibility to recombination may be high in this region.

It is interesting to note that the ancestral sequences to YM29A/YF16 were involved in recombination events that led to the mosaic strains JF6 and the mosaic strains SR631/SR631B. This suggests that the YM29A/YF16 strain played an important role in past evolution of the FIV. Although the SR631/SR631B strains were taken from cougars that are separated from the rest of population by inhospitable terrain, the fact these strains derived ultimately from a recombinant event suggests that in the past the cougars (and hence the strains) were in much closer contact. In particular YM29A/YF16 was an important forebear to the strains that are observed today.

The distant locale of SR631/SR631B compared to other strains suggests that the population structure observed presently within the cougars was different in the past. In particular, the fact that SR631/SR631B was taken from a cougars that are far distant from the rest of the cougars, presents an interesting biogeographic challenge. It suggests that previously the population of cougars from which SR631/SR631B are sampled must have been in geographical proximity to the other cougars, for the mosaic sequences to arise.

FIV from wild cougars has provided a rich source material for which to investigate questions of population structure of the cougars themselves [5] as well as epidemiology of the viruses [6]. Here we suggest that recombination within these strains is interesting in its own right, and as a whole provides a unique opportunity to study recombination in natural populations. We have

addressed the first of many questions, namely which strains appear to be derived from recombinant events and where the recombination seems to be occurring. Many questions remain including understanding the susceptibility to recombination of different genomic regions as well the biogeographic implications of recombination.

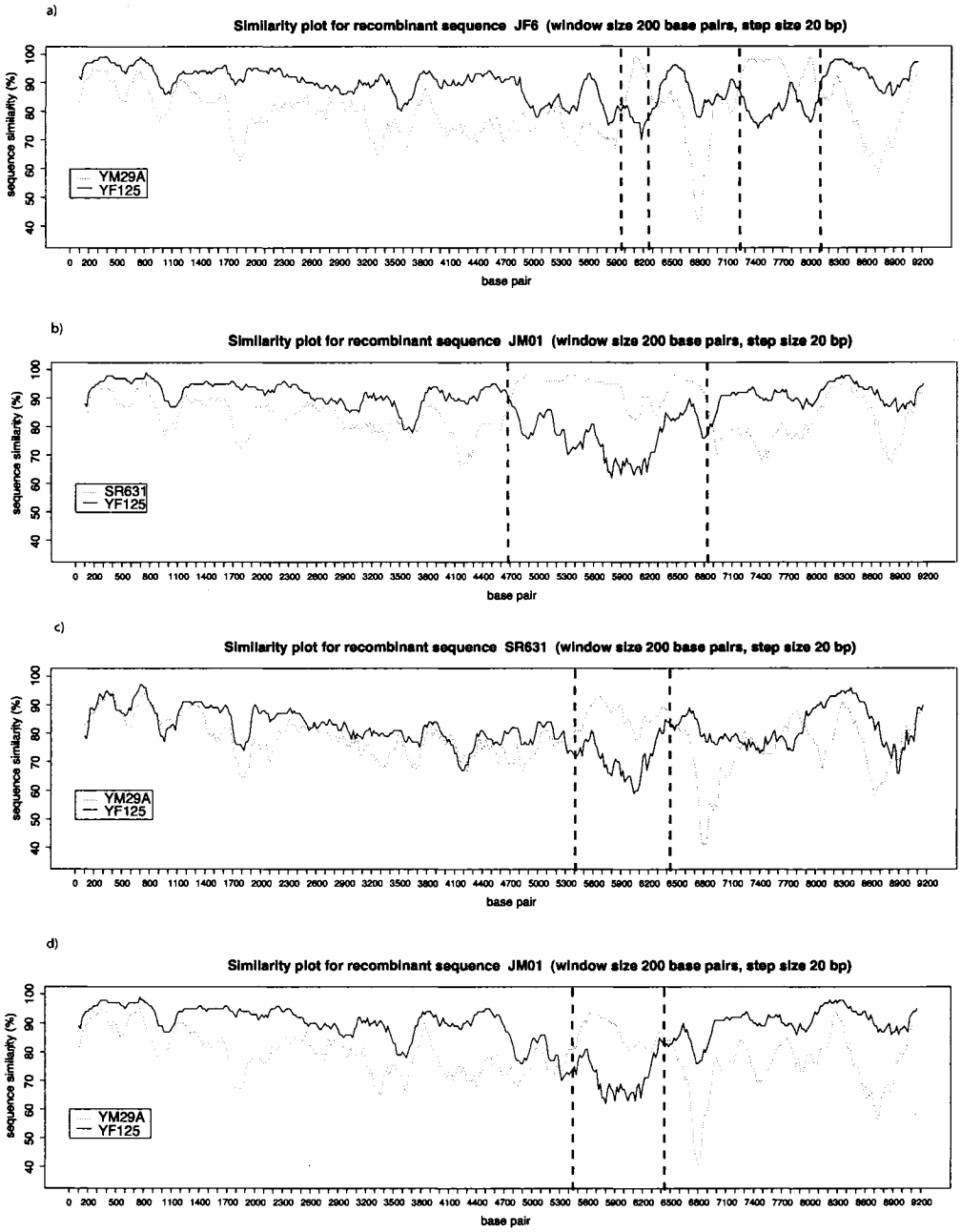


FIGURE 5–4: Major recombinant sequences as shown by similarity plots: a) JF6 with YM29A, b) JM01 with SR631, c) SR631with YM29A and d) JM01 with YM29A. The dashed lines correspond to maximum likelihood estimate of recombinant points and note their correspondence to the areas of significant recombination (shaded regions) in Figure 5–1

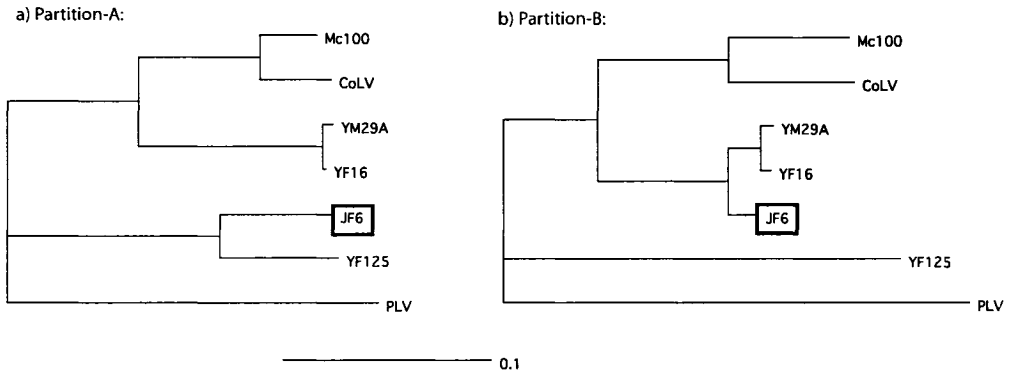


FIGURE 5–5: Maximum likelihood trees (drawn to scale) inferred for a) Partition-A and b) Partition-B. The trees differ in their placement of recombinant sequence JF6. Partitions (including nucleotide substitution models) described in Table 5–1.

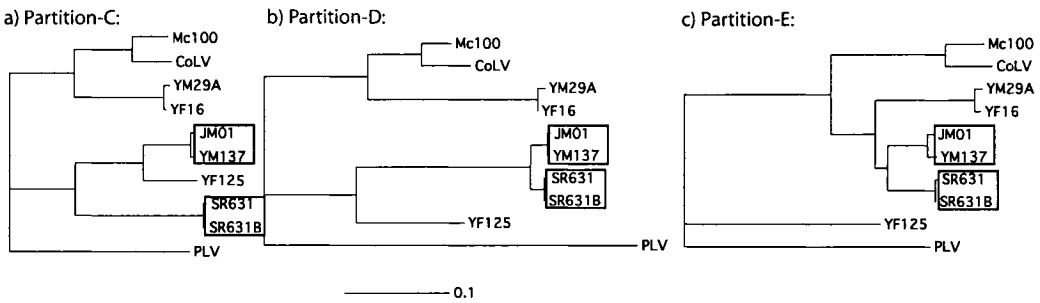


FIGURE 5–6: Maximum likelihood trees (drawn to scale) for partition a) Partition-A, b) Partition-B and c) Partition-C with recombinant sequences JM01/YM137 and SR631/SR631B. Partitions (including nucleotide substitution models) described in Table 5–1.

References

[1] J. P. Anderson, A. G. Rodrigo, G. H. Learn, A. Madan, C. Delahunty, M. Coon, M. Girard, S. Osmanov, L. Hood, and J. I. Mullins. Testing the hypothesis of a recombinant origin of human immunodeficiency virus type 1 subtype e. *Journal of Virology*, 74(22):10752–10765, 2000.

[2] P. Awadalla. The evolutionary genomics of pathogen recombination. *Nature Reviews Genetics*, 4(1):50–60, 2003.

[3] M. H. Bachmann, C. Mathiason-Dubard, G. H. Learn, A. G. Rodrigo, D. L. Sodora, P. Mazzetti, E. A. Hoover, and J. I. Mullins. Genetic diversity of feline immunodeficiency virus: dual infection, recombination, and distinct evolutionary rates among envelope sequence clades. *Journal of Virology*, 71(6):4241–4253, 1997.

[4] M. Bendinelli, M. Pistello, S. Lombardi, A. Poli, C. Garzelli, D. Matteucci, L. Ceccherini-Nelli, G. Malvaldi, and F. Tozzini. Feline immunodeficiency virus: an interesting model for aids studies and an important cat pathogen. *Clinical Microbiology Reviews*, 8(1):87–112, 1995.

[5] R. Biek, A. J. Drummond, and M. Poss. A virus reveals population structure and recent demographic history of its carnivore host. *Science*, 311(5760):538–541, 2006.

[6] R. Biek, A.G. Rodrigo, D. Holley, A. Drummond, C. R. Anderson Jr., H. A. Ross, and M. Poss. Epidemiology, genetic diversity, and evolution of endemic feline immunodeficiency virus in a population of wild cougars. *Journal of Virology*, 77(17):9578–9589, 2003.

[7] E. W. Brown, N. Yuhki, C. Packer, and S. J. O'Brien. A lion lentivirus related to feline immunodeficiency virus: epidemiologic and phylogenetic aspects. *Journal of Virology*, 68(9):5953–5968, 1994.

[8] T. C. Bruen, H. Philippe, and D. Bryant. A simple and robust statistical test for detecting recombination. *Genetics*, 172:2665–2681, 2006.

[9] M. A. Carpenter and S. J. O'Brien. Coadaptation and immunodeficiency virus: lessons from the felidae. *Current Opinion in Genetics and Development*, 5(6):739–745, 1995.

[10] J. Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[11] R. Galetto and M. Negroni. Mechanistic features of recombination in hiv. *AIDS Reviews*, 7:92–102, 2005.

[12] G. Hamilton. Virology: The gene weavers. *Nature*, 441(7094):683–685, 2006.

[13] E. C. Holmes, M. Worobey, and A. Rambaut. Phylogenetic evidence for recombination in dengue virus. *Molecular Biology and Evolution*, 16(3):405–9, 1999.

[14] G. Hommel. A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, 75(2):383–386, 1988.

[15] C. M. Johnson, B. A. Torres, H. Koyama, and J. K. Ymamoto. Tenth anniversary perspectives on aids. fiv as a model for aids vaccination. *Aids Research and Human Retroviruses*, 10(3):225–228, 1994.

[16] K. S. Lole, R. C. Bollinger, R. S. Paranjape, D. Gadkari, S. S. Kulkarni, N. G. Novak, R. Ingersoll, H. W. Sheppard, and S. C. Ray. Full-length human immunodeficiency virus type 1 genomes from subtype c-infected seroconverters in india, with evidence of intersubtype recombination. *Journal of Virology*, 73(1):152–60, 1999.

[17] J. Maynard Smith. Analyzing the mosaic structure of genes. *Journal of Molecular Evolution*, 34:126–9, 1992.

[18] R. A. Olmsted, R. Langley, M. E. Roelke, R. M. Goeken, D. Adger-Johnson, J. P. Goff, J. P. Albert, C. Packer, M. K. Laurenson, and T. M. Caro. Worldwide prevalence of lentivirus infection in wild feline species: epidemiologic and phylogenetic aspects. *Journal of Virology*, 66(10):6008–6018, 1992.

[19] R. D. M. Page. Treeview: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences*, 12:357–58, 1996.

[20] N. C. Pedersen, E. W. Ho, M. L. Brown, and J. K. Yamamoto. Isolation of a T-lymphotropic virus from domestics cats with an immunodeficiency-like syndrome. *Science*, 235(4790):790–793, 1987.

[21] D. Posada. Evaluation of methods for detecting recombination from DNA sequences: Empirical data. *Molecular Biology and Evolution*, 19(5):708–717, 2002.

[22] D. Posada and K. A. Crandall. MODELTEST: testing the model of DNA substitution. *Bioinformatics*, 14(9):817–818, 1998.

[23] D. Posada and K. A. Crandall. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *PNAS*, 98(24):13757–13762, 2001.

[24] D. Posada, K. A. Crandall, and E. C. Holmes. Recombination in evolutionary genomics. *Annual Review of Genetics*, 36(1):75–97, 2002.

[25] F. Reggeti and D. Bienzle. Feline immunodeficiency virus subtypes a, b and c and intersubtype recombinants in ontario, canada. *Journal of General Virology*, 85(7):1843–1852, 2004.

[26] D. L. Robertson, P. M. Sharp, F. E. McCutchan, and B. H. Hahn. Recombination in HIV-1. *Nature*, 374(6518):124–126, 1995.

[27] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

[28] M. Salemi, T. De Oliveira, V. Courgnaud, V. Moulton, B. Holland, S. Cassol, W. M. Switzer, and A. Vandamme. Mosaic genomes of the six major primate lentivirus lineages revealed by phylogenetic analyses. *Journal of Virology*, 77(13):7202–7213, 2003.

[29] M. O. Salminen, J. K. Carr, D. S. Burke, and F. E. McCutchan. Identification of breakpoints in intergenotypic recombinants of HIV type-1 by bootscanning. *Aids Research and Human Retroviruses*, 11(11):1423–1425, 1995.

[30] M. H. Schierup and J. Hein. Consequences of recombination on traditional phylogenetic analysis. *Genetics*, 156(2):879–891, 2000.

[31] H. Shimodaira and M. Hasegawa. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, 16:1114–1116, 1999.

[32] D. L. Swofford. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods).* Version 4. Sinauer Associates, Sunderland, Massachusetts, 2001.

[33] T. Uzzell and K.W. Corbin. Fitting discrete probability distributions to evolutionary events. *Science*, 172:1089–1096, 1971.

[34] C. Wiuf, T. Christensen, and J. Hein. A simulation study of the reliability of recombination detection methods. *Mol Biol Evol*, 18(10):1929–1939, 2001.

[35] Z. Yang. Maximum-likelihood estimation of phylogeny from dna sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, 10(6):1396–1401, 1993.

CHAPTER 6

Summary and further questions

6.1 Summary and further questions

A number of different topics were touched on in this thesis, relating from mathematical phylogenetics, to methods for detecting recombination and finally to understanding the evolution of FIV. A number of natural questions arise from each of the chapters, which we discuss here.

6.1.1 Mathematical phylogenetics

Chapter 2 showed how maximum parsimony was related to maximum compatibility through a notion of character subdivision. The relationship was exploited to give a simple formula to calculate the parsimony score for a pair of characters. It is well known however, that calculating the parsimony score for an arbitrary number of characters is NP-complete [2]. Thus a natural question is to characterize the complexity for three or more characters. This has already been done for the character compatibility problem [1] and the close relationship between this problem and maximum parsimony problem suggests that this is a realistic question to address.

Further graph theoretic consequences of the result may be obtained using the relationship between vertex subdivision an maximum parsimony.

Chapter 3 connects different forms of parsimony with a metric space on the set of leaf-labeled trees under common tree metrics (such as SPR and RF). The work may be applicable to help visualize tree space [3], by determining the increase in parsimony score after tree re-arrangement operations (personal communication Katherine St. John). This result also complements recent algorithmic approaches in phylogenetic networks used for simplifying networks (personal communication, Mike Steel). Interesting questions remain such as exploring analogous connections with tree space under a probabilistic model of mutation.

6.1.2 Statistically methods for understanding recombination

Chapter 4 uses the formula for calculating the parsimony score for two characters to develop a statistic for detecting recombination. Chapter 4 shows that the new method performs as well if not better than comparable methods. However, Chapter 4 shows that all methods fail at distinguishing recombination from recurrent mutation when there are very few mutations. This has practical implications since mitochondrial DNA has very few mutations and thus distinguishing hypervariable sites from recombination is challenging. Fundamentally Chapter 4 relies on the concept of compatibility which for smaller population mutation rates, θ leaves most recombinant events undetectable [4]. Thus an important question is to address this issue,

that is determine methods that are sensitive to very little mutation (but still robust against false positives).

6.1.3 Recombination in FIV

Chapter 5 applied the statistical test developed in Chapter 4 to explore the evolution of FIV in a wild population of cougars. Since the test in Chapter 4 simply determines whether recombination is present or not, this test was applied multiple times to different genomic locations. By applying the test multiple times regions containing recombinant breakpoints were identified, leading ultimately to the identification of recombinant strains. The procedure was very conservative however since multiple test correction was used and the dependency between tests was not fully exploited. Moreover, multiple steps were needed to go from breakpoint area identification to recombinant strain identification. Preferably, these steps could be automated using a compatibility approach, with statistical uncertainty properly assessed.

References

[1] H. Bodlaender, M. Fellows, M. Hallett, T. H. Wareham, and T. Warnow. The hardness of Perfect Phylogeny, Feasible Register Assignment and other problems on thin colored graphs. *Journal of Theoretical Computer Science*, 244:167–188, 2000.

[2] L. R. Foulds and R. L. Graham. The Steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics*, 3:43–49, 1982.

[3] D. M. Hillis, T. Heath, and K. St. John. Analysis and visualization of tree space. *Systematic Biology*, 54:471–482, 2005.

[4] R. R. Hudson and N. L. Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111(1):147–64, 1985.

Appendix

Permission to include published papers in PhD Thesis

Paper:

Bruen, T. C., Philippe, H. and Bryant D. (2006) A simple and robust statistical test for detecting recombination. *Genetics.* 172(6):2665-2681.

(See next page)

GENETICS
Box I
Mellon Institute
4400 Fifth Avenue
Pittsburgh PA 15213

Dear Dr. Bruen:

In response to your request, you have the permission of the Genetics Society of America to reprint your article, including text, figures, and legends, in your disseratation. The article was published as:

Genetics, Vol. 172, 2665-2681, April 2006, Copyright © 2006
doi:10.1534/genetics.105.048975

A Simple and Robust Statistical Test for Detecting the Presence of Recombination

Trevor C. Bruen, Hervé Philippe and David Bryant

This material may be deposited in the Library and Archives of Canada (formerly National Library of Canada).

Many thanks,
tracey

Tracey DePellegrin Connelly
Managing Editor
GENETICS

DOI: 10.1534/genetics.105.048975

A Simple and Robust Statistical Test for Detecting the Presence of Recombination

Trevor C. Bruen,*,[1] Hervé Philippe[†] and David Bryant*,[‡]

**McGill Centre for Bioinformatics, McGill University, Montreal, Quebec H3A 2B4, Canada, [†]Program in Evolutionary Biology, Canadian Institute for Advanced Research, Centre Robert Cedergren, Département de Biochimie, Université de Montréal, Montreal, Quebec H3T 1J4, Canada and [‡]Department of Mathematics, University of Auckland, Auckland, New Zealand*

Manuscript received July 30, 2005
Accepted for publication February 3, 2006

ABSTRACT

Recombination is a powerful evolutionary force that merges historically distinct genotypes. But the extent of recombination within many organisms is unknown, and even determining its presence within a set of homologous sequences is a difficult question. Here we develop a new statistic, Φ_w that can be used to test for recombination. We show through simulation that our test can discriminate effectively between the presence and absence of recombination, even in diverse situations such as exponential growth (star-like topologies) and patterns of substitution rate correlation. A number of other tests, Max χ^2, NSS, a coalescent-based likelihood permutation test (from LDHat), and correlation of linkage disequilibrium (both r^2 and $|D'|$) with distance, all tend to underestimate the presence of recombination under strong population growth. Moreover, both Max χ^2 and NSS falsely infer the presence of recombination under a simple model of mutation rate correlation. Results on empirical data show that our test can be used to detect recombination between closely as well as distantly related samples, regardless of the suspected rate of recombination. The results suggest that Φ_w is one of the best approaches to distinguish recurrent mutation from recombination in a wide variety of circumstances.

RECOMBINATION is a fundamental biological process that can, for example, increase viral or bacterial pathogenicity by diffusing genetic material throughout populations (Awadalla 2003). The biological mechanisms of recombination differ across organisms, but in broad terms recombination results in the creation of mosaic sequences where the evolutionary history at each site may be different. Violating this tree-like assumption of evolution can lead to serious consequences when performing phylogenetic analyses for a set of sequences. Indeed, as the evolution of the sequences cannot be described by a single tree, this can lead to overestimation or underestimation of branch lengths among other problems (Schierup and Hein 2000a,b; Posada 2001; Posada and Crandall 2002). Thus, an important question for a given set of aligned sequences is to determine whether or not recombination is likely to have occurred.

The ability of a large number of general methods to detect recombination has recently been evaluated empirically and through simulation (Crandall and Templeton 1999; Brown *et al.* 2001; Posada and Crandall 2001; Wiuf *et al.* 2001; Posada 2002). These studies have established that methods such as Geneconv (Sawyer 1989), Max χ^2 (Maynard Smith 1992), RDP (Martin and Rybicki 2000), Phypro (Weiller 1998), RecPars (Hein 1990, 1993), and neighbor similarity score (NSS) (Jakobsen and Easteal 1996) efficiently detect recombination in a wide range of circumstances (Brown *et al.* 2001; Posada and Crandall 2001; Wiuf *et al.* 2001; Posada 2002). These tests infer the presence of recombination either directly through sequence comparisons or indirectly through phylogenetic means. As no underlying assumptions are made concerning the origin of the sequences, these tests can be applied to detect recombination within any set of aligned homologous sequences. Indeed, these techniques can be used to detect recombination within either closely or distantly related genotypes (Posada 2002). Moreover, these methods can be termed general since no specific assumptions concerning sample history (beyond sequence homology) are made.

In contrast to general methods for inferring recombination, there are also population-specific methods for detecting recombination, where the samples consist of genotypes from closely related individuals. Within a single population, recombination can be tested for using nonparametric approaches such as permutation tests based on summary statistics like the correlation of linkage disequilibrium with distance (Miyashita and Langley 1988; Schaeffer and Miller 1993; Awadalla *et al.* 1999). Linkage disequilibrium is typically measured

[1]*Corresponding author:* McGill Centre for Bioinformatics, Duff Medical Bldg., 3775 University St., Montreal, QC H3A 2B4, Canada.
E-mail: trevor@mcb.mcgill.ca

using the statistics r^2 and $|D'|$ (Lewontin 1964; Hill and Robertson 1968).

Recently, coalescent (Kingman 1982) methods have been developed that can specifically detect (Brown *et al.* 2001; McVean *et al.* 2002) or characterize the rate of recombination (Griffiths and Marjoram 1996; Hey and Wakeley 1997; Kuhner *et al.* 2000; Nielsen 2000; Wall 2000; Fearnhead and Donnelly 2001; Hudson 2001; McVean *et al.* 2002) for a set of samples within a single population. Recombination can be modeled under either a basic crossing-over model (Hudson 1983) or a more complex model of gene conversion (Wiuf and Hein 2000). Only a few methods (Kuhner *et al.* 2000; Fearnhead and Donnelly 2001; McVean *et al.* 2002) relax the infinite-sites model (Kimura 1969) under which a site can undergo at most a single mutation. Relaxing the infinite-sites model is important for many bacterial and viral data sets, since under the infinite-sites model, high levels of recurrent mutation can cause patterns consistent with recombination (McVean *et al.* 2002).

The basic coalescent operates under several assumptions that include constant population size, no selection, random mating, and no population structure (Hein *et al.* 2005). Whereas these assumptions can be relaxed using additional parameters such as a term for population growth (Slatkin and Hudson 1991), these additional parameters are presently not accounted for in current methods that characterize and detect recombination (Kuhner *et al.* 2000; Fearnhead and Donnelly 2001; McVean *et al.* 2002). Importantly, the influence of population structure and demographic history may adversely affect the ability of coalescent methods to correctly infer the rate of recombination (McVean *et al.* 2002; Haydon *et al.* 2004).

The myriad of methods available to detect, characterize, and find recombinant sequences is somewhat bewildering. Traditionally, general approaches have been used for recombination analysis between distantly related genotypes, whereas population genetic-based approaches have been used for recombination analysis between closely related genotypes. However, in many cases the line between the approaches is blurred, and both approaches have been used to infer the presence of recombination in bacteria, viral, and animal mitochondrial data sets (McVean *et al.* 2002; Posada 2002; Piganeau *et al.* 2004).

Often, one of the primary questions for any data analysis is to determine whether recombination is likely to be present within a set of sequences at all (Awadalla *et al.* 1999; Maynard Smith and Smith 2002; McVean *et al.* 2002; Posada 2002; Piganeau *et al.* 2004; Tsaousis *et al.* 2005). Indeed, there are still open questions with regard to the extent of recombination in animal mitochondrial DNA (Maynard Smith and Smith 2002; Piganeau *et al.* 2004; Tsaousis *et al.* 2005). Moreover, if the sequences are obtained from closely related, yet distinct, organisms or from many different populations, it is inappropriate to analyze the sequences in a framework that assumes a single population, such as linkage disequilibrium or coalescent approaches (Tsaousis *et al.* 2005). But determining whether recombination has occurred in such circumstances is an important question that cannot be easily answered in a parametric framework. A robust nonparametric test for recombination can help distinguish between the presence and absence of recombination in such cases.

Testing for recombination can statistically validate visual evidence of recombination obtained using, for instance, phylogenetic network approaches (*e.g.*, Huson and Bryant 2006) or independently verify the presence of recombination if a positive estimate of the rate of recombination is inferred (*e.g.*, McVean *et al.* 2002). Moreover, it is often difficult to distinguish between rate heterogeneity and recombination in many circumstances (Grassly and Holmes 1997; McGuire and Wright 2000) and thus regions that exhibit phylogenetic inconsistencies can be individually tested for recombination. Additionally, testing for recombination can be used as a prior probability for the presence of recombination when inferring the points at which infrequent recombination may have occurred (Minin *et al.* 2005). In this sense, testing for recombination can be used in conjunction with other methods.

Ideally, a single test could correctly determine whether recombination is present within any given set of aligned sequences, regardless of population history, demographic history, recombination rate, or mutation rate. Preferably, such a test would also minimize the production of false positives. Here we develop a new test that is powerful under many of these different situations and produces few false positives. Through simulation and empirical data analysis we characterize the performance of our test under various rates of recombination, rates of mutation, demographic histories, and sample sizes. We also show through simulation that a simple model of substitution rate autocorrelation (consistent with mutational "hot spots") gives rise to a signal similar to recombination for two different general tests, Max χ^2 and NSS, but not for our method.

METHODS

Tests for recombination based on the principle of compatibility have proved to be among the most powerful (Brown *et al.* 2001; Posada and Crandall 2001; Wiuf *et al.* 2001; Posada 2002). The traditional binary notion of compatibility (Le Quesne 1969) is well suited for sites with at most two alleles, but can be directly extended into a broader notion (Penny and Hendy 1986) that we term here as refined incompatibility. We then develop a new statistic to test for recombination, the Φ_w (or pairwise homoplasy index, PHI) statistic that uses this notion of refined incompatibility.

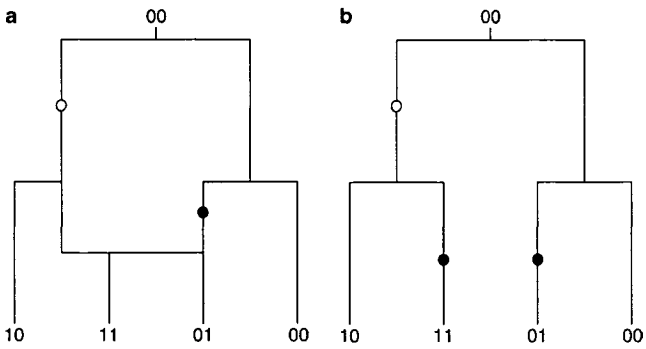


FIGURE 1.—The dual nature of incompatibility. Two possible histories for a pair of incompatible sites are shown: (a) two incompatible sites explained by a recombination event and (b) two incompatible sites explained by a convergent mutation. Mutations in the first site are indicated by open circles and mutations in the second site are indicated by solid circles. To explain the incompatibility between the pair of sites either a recombination event must be invoked or a homoplasy must have occurred in the history of one of the sites.

Compatibility and incompatibility: It is not obvious how to determine the genealogical history of a single site. As such, the pattern of mutation present at multiple sites must be used to infer the genealogy of the sample as a whole. One possibility is to use the observed patterns at pairs of sites, in particular the notion of compatibility (LE QUESNE 1969) or the "four-gametes" test (HUDSON and KAPLAN 1985). Two sites i and j are compatible if and only if there is a genealogical history that can be inferred parsimoniously that does not involve any recurrent or convergent mutations (known as homoplasies as in Figure 1b). If the two sites are not compatible, they are termed incompatible. Under an infinite-sites model (KIMURA 1969) of sequence evolution, the possibility of a homoplasy does not exist, and so incompatibility for a pair of sites implies that at least one recombination event must have occurred, as in Figure 1a. This can be used to estimate the minimum number of recombination events present in the sample as a whole (HUDSON and KAPLAN 1985; SONG AND HEIN 1999; MYERS and GRIFFITHS 2003). Testing for compatibility can be accomplished by checking if all four combinations of {00, 01, 10, 11} are present among the sequences (LE QUESNE 1969).

The traditional, binary notion of either compatibility or incompatibility treats a single homoplasy the same as many homoplasies. That is, although in some situations more than one homoplasy can be parsimoniously inferred for a pair of sites (CAMIN and SOKAL 1965; PENNY and HENDY 1986), this information is disregarded. Consider two sites i and j, with $|\chi_i|$ and $|\chi_j|$ representing the number of observed states (alleles) at each site. Let $l(\chi_i, \chi_j)$ denote the minimum number of mutations required by *any tree* used to represent the genealogical history of both sites. Thus $l(\chi_i, \chi_j)$ represents the maximum parsimony score for these two characters over all trees. Note that $l(\chi_i, \chi_j) \geq (|\chi_i| - 1) + (|\chi_j| - 1)$ as each state (except the ancestral state) must arise at least once in the tree. Define the refined incompatibility score of sites i and j as

$$i(\chi_i, \chi_j) = l(\chi_i, \chi_j) - (|\chi_i| - 1) - (|\chi_j| - 1).$$

The refined incompatibility score relates to the traditional notion of compatibility in the following way: two sites are compatible if and only if $i(\chi_i, \chi_j) = 0$; if $i(\chi_i, \chi_j) > 0$ the two sites are incompatible. There are also two interpretations of this refined incompatibility score: in the absence of recombination, this score represents the minimum number of homoplasies that have occurred in the history of the samples for these two sites (PENNY and HENDY 1986); in the absence of recurrent or convergent mutations, this score represents the minimum number of recombinations that have occurred between the two sites (T. BRUEN and D. BRYANT, unpublished data). This latter result depends on viewing recombinations as unrooted subtree-prune and regraft operations (see HEIN *et al.* 2005). Importantly, this score can be calculated quickly [linear time in the number of sequences (BRUEN and BRYANT 2006)], which allows alignments with large numbers of sequences to be evaluated rapidly.

A parsimony informative site has at least two different alleles that are represented by at least two different sequences each (there must be at least four sequences at a site for the site to be parsimony informative) (FELSENSTEIN 2004). A compatibility matrix (SNEATH *et al.* 1975; JAKOBSEN and EASTEAL 1996) is traditionally used to represent compatibility between all pairs of parsimony informative sites. This matrix can also easily be extended into a refined incompatibility matrix by setting each entry (i, j) equal to the refined incompatibility score between any two sites i and j.

Sites that have the same history will tend to be more compatible than sites that have different histories (SNEATH *et al.* 1975; JAKOBSEN and EASTEAL 1996; DROUIN *et al.* 1999). One way to measure the extent of "clustering" in the matrix is to consider the proportion of neighboring cells in the matrix that are either compatible or incompatible. The resulting statistic is termed the NSS and has been used as a powerful test for recombination (JAKOBSEN and EASTEAL 1996; BROWN *et al.* 2001; POSADA and CRANDALL 2001; WIUF *et al.* 2001; POSADA 2002). However, simulations suggest that the NSS produces an excess of "false positives" in certain situations (see RESULTS AND DISCUSSION) and so we have developed an alternative statistic.

Test statistic (Φ_w): The degree of genealogical correlation between neighboring sites is negatively correlated with the rate of recombination (HUDSON and KAPLAN 1985). In the case of finite levels of recombination, the genealogical correlation of sites is partially reflected by a tendency of closely linked sites to have greater compatibility than distant sites (HAGENBLAD and NORDBORG 2002; INNAN and NORDBORG 2002).

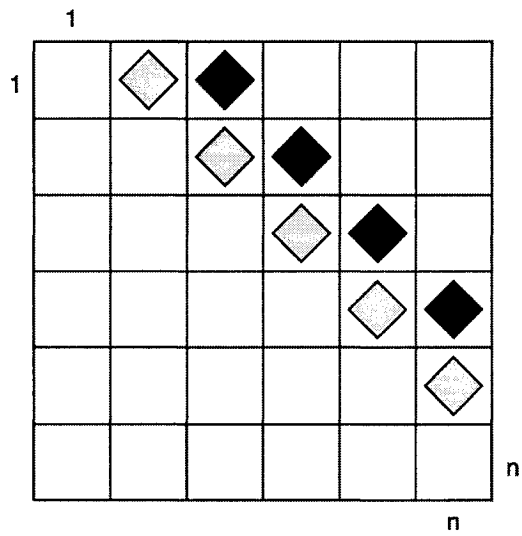


FIGURE 2.—The entries marked with a diamond in the refined incompatibility matrix represent the cells used to calculate the pairwise homoplasy index (or Φ_w). The cells with light shading contain the refined incompatibility score of informative site i with informative site $i + 1$. The cells with dark shading contain the refined incompatibility score of informative site i with informative site $i + 2$. In this example sites up to 2 informative bases apart are used to calculate Φ_w.

To measure the similarity between closely linked sites, we propose calculating a new statistic, the pairwise homoplasy index (PHI). The idea is to calculate the mean refined incompatibility score from nearby sites by using the first k off-diagonal rows of a refined incompatibility matrix (see Figure 2). Let w denote a fixed width (measured in bases) and choose k so that it is proportional to w. Specifically, let q denote the proportion of parsimony informative sites within the alignment and set $k = wq$. The statistic thus measures the mean refined incompatibility score of sites up to (approximately) w bases apart. We can now formally define the Φ or PHI statistic as

$$\Phi_w = \frac{2}{k(2n-k-1)} \sum_{j=1}^{k} \sum_{i=1}^{n-j} i(\chi_i, \chi_{i+j}).$$

The term "pairwise homoplasy index" refers to the fact that the refined incompatibility score can be interpreted as the minimum number of convergent or recurrent mutations (homoplasies) necessarily present on any tree describing the history of any two sites i and j. The term $k(2n - k - 1)/2$ is a normalizing factor.

Clearly w should be somewhat less than the total number of sites but large enough that a number of comparisons are made. For all simulated and empirical analyses w was set to 100 and k chosen according to the above formula. Other choices of w were also considered ($w = 50$ and $w = 150$), but simulations (across different sequence lengths) suggested that $w = 100$ was slightly better than the other two choices (results not shown).

Significance: Significance of the observed Φ_w-statistic can be obtained by using a permutation test. Under the null hypothesis of no recombination, the genealogical correlation of adjacent sites is invariant to permutations of the sites as all sites have the same history. But in the case of finite levels of recombination, the order of the sites is important, as distant sites will tend to have less genealogical correlation than adjacent sites. Let $\hat{z}$ denote the observed value of the Φ_w-statistic on the original alignment and let Z_0 denote the value of the Φ_w-statistic for a random permutation of the sites. Hence Z_0 is distributed according to the null hypothesis of no recombination. To determine the significance of the observed value $\hat{z}$, a Monte Carlo P-value can be directly estimated by permuting the alignment many times and counting the proportion of times the Φ_w-statistic on a permuted alignment is less than or equal to $\hat{z}$. However, computation of P-values based on permutations of the alignment is time consuming. One way to circumvent this problem is to determine the distribution of the test statistic under permutations of the alignment. The expectation ($E_0(\Phi_w) = \mu'$) and variance ($\mathrm{Var}_0(\Phi_w) = \sigma^2$) of Φ_w can be calculated analytically (see APPENDIX A for details). Moreover, initial simulations indicated that the distribution of Φ_w under permutations of the alignment is approximately normal (results not shown). Using these assumptions, the value of $\Pr(Z_0 \leq \hat{z})$ can be calculated as

$$\Pr(Z_0 \leq \hat{z}) = \int_{-\infty}^{\hat{z}} n(\tau \mid \mu', \sigma^2) d\tau,$$

where $n(\tau \mid \mu', \sigma^2)$ denotes a normal probability distribution function with mean μ' and variance σ^2. This alternative to the permutation test has the advantage that it can be obtained quickly and gives a more precise P-value under an assumption of normality.

The normality of the distribution of the test statistic can be explained by noting that for a large refined incompatibility matrix, calculating the Φ_w-statistic amounts to taking the mean of a small sample of values from the matrix. The simplest version of the central limit theorem then suggests that taking the mean of a small sample within a "large" matrix has a limiting normal distribution, if the terms are independent and identically distributed (CASELLA and BERGER 2001). However, in this case the central limit theorem provides a guide rather than a formal equivalence.

For every data set examined (both simulated and empirical) the significance of the observed Φ_w-statistic was calculated using the permutation test directly as well as the normal alternative. The P-values obtained by using the permutation test are written as $P_P(\Phi_w)$ whereas the P-values obtained by using the normal alternative are written as $P_N(\Phi_w)$.

Simulation study: We repeated many of the same simulations that had been performed in other studies (POSADA and CRANDALL 2001; WIUF *et al.* 2001) but expanded the parameter search space and considered the Φ_w-statistic as well as additional tests. The protocol followed was

based on simulations from the neutral coalescent model (KINGMAN 1982) with recombination (HUDSON 1983).

The coalescent model provides a natural foundation for simulation (CRANDALL and TEMPLETON 1999; BROWN *et al.* 2001; POSADA and CRANDALL 2001; WIUF *et al.* 2001). Simulations were almost all conducted using the program Treevolve (GRASSLY *et al.* 1999). For very high rates of recombination ($\rho = 128$), simulations were performed using the program Hudson (SCHIERUP and HEIN 2000a,b) since the program Treevolve did not run at such high rates of recombination. Mutations were added according to a Jukes–Cantor model (JUKES and CANTOR 1969). Other methods of sequence evolution were also examined, including the addition of extreme rate heterogeneity ($\alpha = 0.1$), which resulted in a moderate decrease in power for all methods (results not shown). For each parameter setting, 1000 replicate data sets were created, with each replicate consisting of an alignment of length 1000 (see APPENDIX B for further details). Significance was set at the 0.05 level.

In addition to the Φ_w-statistic, four of the best nonparametric tests were computed for each parameter setting, namely the Max χ^2-statistic (MAYNARD SMITH 1992), the NSS (JAKOBSEN and EASTEAL 1996), and two measures of correlation of linkage disequilibrium (r^2 and $|D'|$) with distance (LEWONTIN 1964; HILL and ROBERTSON 1968; MIYASHITA and LANGLEY 1988; SCHAEFFER and MILLER 1993). Furthermore, results obtained from a coalescent-based likelihood permutation test (LPT) from LDHat (MCVEAN *et al.* 2002) are reported as well. The Max χ^2-statistic has been found to be the best general test for detecting recombination in a recent empirical study (POSADA 2002), and the NSS statistic has been found to be very efficient as well (BROWN *et al.* 2001; POSADA and CRANDALL 2001; WIUF *et al.* 2001; POSADA 2002). Correlation of linkage disequilibrium with distance using r^2 has been found to be the strongest nonparametric approach for detecting recombination within populations (MCVEAN *et al.* 2002). Recently, the likelihood permutation test was introduced as a powerful alternative to methods based on linkage disequilibrium (MCVEAN *et al.* 2002). For the Max χ^2-statistic a fixed window size of the number of polymorphic sites divided by 1.5 was used following a previously described protocol (POSADA and CRANDALL 2001; POSADA 2002). For both measures of correlation of r^2 and D' with distance, only sites with two alleles segregating and minor allele frequencies of at least 0.1 were used, as this approach tends to maximize power (WEIR and HILL 1986; MCVEAN *et al.* 2002). For the likelihood permutation test, precomputed likelihood files were used on the basis of 101 grid points with a value of θ per site of either 0.001 or 0.1. For each replicate, if the expected mean sequence diversity was $<10\%$, then a likelihood file with a θ per site value of 0.001 was used; otherwise a likelihood file with a θ per site value of 0.1 was used (under a constant-size population the expected mean sequence diversity of 10% corresponds to an expected value of θ per site of ~ 0.12). The significance for each of the statistics was obtained using a permutation test. For the power determination, 1000 permutations were performed, whereas for the false positives, 200 permutations were performed.

Power: To determine power in the presence of recombination, the recombination rate ρ (under population growth $\rho^\dagger$) varied among 0, 1, 2, 4, 8, 16, and 128; the expected nucleotide diversity p between any two sequences varied among 1, 5, 10, 15, and 25%; and the growth rate of the population β varied between 0 (constant-size populations) and 5000. The sample size m varied among 5, 10, 15, 25, and 50. For $\rho = 128$ simulations with $\beta = 5000$ were not performed since this option was not available with the program Hudson. More details explaining the protocol can be found in APPENDIX B and elsewhere (WIUF *et al.* 2001).

False positives: Substitution rate heterogeneity across sites on a genealogy was modeled here using a Γ-distribution (UZZELL and CORBIN 1971; YANG 1993). In this case, the substitution rate at each site i, Z_i, is drawn from a Γ-distribution with shape parameter α and scale parameter $1/\alpha$ (YANG 1993).

Autocorrelation among substitution rates was modeled assuming Markov dependence among rates (YANG 1995). To achieve this, two random variables Y_i and Y_{i+1} were drawn from a bivariate normal distribution with correlation ρ_N and transformed into two marginally distributed gamma random variables Z_i and Z_{i+1} with correlation ρ_G (YANG 1995). Using the bivariate normal distribution of Y_i and Y_{i+1} (including correlation ρ_N), the probability distribution function of random variable Y_{i+1} was obtained conditional on the random variable Y_i, allowing Markov-dependent substitution rates to be drawn. The substitution rates Z_i and Z_{i+1} then represent draws from a bivariate Γ-distribution with correlation ρ_G. The value of ρ_G is positively correlated with the value ρ_N but not identical (YANG 1995).

Data sets were simulated using a modified version of Treevolve (GRASSLY *et al.* 1999) with a number of the sampling functions taken from PAML (YANG 1997). The correlation parameter ρ_N varied among 0 (no correlation), 0.3, 0.6, and 0.9; the expected nucleotide diversity p between any two sequences varied among 1, 5, 10, 15, and 25%; the value of α for the Γ-distribution varied among 0.1, 1.0, and ∞; and the growth rate of the population β varied between 0 (constant-size populations) and 5000. The sample size m varied among 5, 10, 15, 25, and 50.

Empirical data: A number of population and species level data sets were examined. The presence of recombination in each of these data sets was debated, unknown, or suspected. The rate of recombination in these data sets ranged from rare to very frequent. In general, data sets with at least a few hundred sites were chosen.

Tests for recombination were performed using the Φ_w-statistic as well as the Max χ^2-statistic (MAYNARD SMITH

TABLE 1

Summary of empirical data sets

Data set	Type	No. of sequences	No. of sites	Informative sites	Observed diversity (%)[a]	Tajima's D[b]	Reference
Candida albicans	Fungi	45	2553	58	0.7	0.936	Anderson *et al.* (2001)
Rana	Animal mtDNA	8	1143	257	14.8	—	Sumida *et al.* (2000)
Cowdria ruminantium	Bacteria	14	870	186	10.5	0.384	Jiggins (2002)
H. pylori	Bacteria	33	472	53	3.8	−0.531	Suerbaum *et al.* (1998)
Boletales	Fungi	31	639	265	17.1	—	Kretzer and Bruns (1999)
Norovirus	Virus	25	1617	103	2.2	−1.482	Rohayem *et al.* (2005)
Apodemus	Animal mtDNA	10	1140	275	14.7	—	Martin *et al.* (2000)
Nematode Wolbachia	Bacteria	10	444	98	13.0	0.899	Jiggins (2002)

[a] Mean proportion of sites that differ between any two sequences.
[b] Calculated on sites with only two alleles segregating.

1992) and the NSS statistic (Jakobsen and Easteal 1996). As in the simulation studies, w was set to 100 for all analyses. One thousand permutations were performed to obtain significance. Additional results are reported for the population level data sets, using permutation tests based on r^2 and $|D'|$ (Lewontin 1964; Hill and Robertson 1968; Miyashita and Langley 1988; Schaeffer and Miller 1993) as well as a coalescent-based LPT with LDHat (McVean *et al.* 2002). Furthermore, an estimate of the rate of recombination was also obtained in LDHat using a model of crossing over rather than gene conversion. The maximum value of ρ was set to 100 and 100 grid points were used in LDHat. The value of Tajima's D-statistic is also reported, as it can be an indicator of population growth or selective pressure (Tajima 1989). Table 1 summarizes the data sets used. The data sets include sequences from bacteria, viruses, and fungi. Two of the data sets were from animal mitochondrial DNA (mtDNA).

For the Boletales data set additional analysis was performed by first estimating a neighbor-joining tree (Saitou and Nei 1987) using PAUP* (Swofford 1998). Branch lengths for the tree, a transition/transversion ratio, codon frequencies, a value of α for the substitution rate heterogeneity (Yang 1993), as well as the degree of substitution rate autocorrelation (estimated using the autodiscrete gamma model) (Yang 1995), were then estimated using a codon model in PAML (Yang 1997). A parametric bootstrap of 1000 replicates was then performed under the estimated parameters using a modified version of PAML that allowed autocorrelated substitution rates. For each replicate, a test for recombination was performed using the Max χ^2-statistic, the NSS statistic, and the Φ_w-statistic (with 1000 permutations). Significance was set at 0.05.

RESULTS AND DISCUSSION

Simulation studies: *Analytical calculation of P-values:* Table 2 shows the proportion of times that recombination was inferred using Φ_w, when the rate of recombination ρ was set to 0 and there was no population growth ($\beta = 0$). Since the significance level was set to 0.05, the Φ_w-test is too conservative when the mean sequence diversity is ~1% or when there are few samples (*e.g.*, $m = 5$). This is partly due to the fact that there are very few informative sites or incompatibilities produced in these situations (results not shown). Table 2 also indicates that when the sequence diversity and sample size are small, obtaining significance using the permutation test ($P_P(\Phi_w)$) is even more conservative than obtaining significance using the normal distribution ($P_N(\Phi_w)$). On the other hand, Figure 3 shows that both methods for obtaining significance give very similar answers for higher amounts of sequence diversity (at least 10%), with at least 15 samples. These results suggest that it is sufficient to obtain significance for Φ_w using the normal distribution. For all subsequent simulations, the results quickly obtained with the Φ_w-statistic using the normal distribution are reported.

Time: The time to calculate Φ_w is much faster than other population genetic methods especially for moderate numbers of sites and sequences. For instance, several simulated alignments of 25 samples with 5000 sites with moderate sequence diversity (10%), corresponding

TABLE 2

Proportion of times recombination inferred using Φ_w when $\rho = 0$ and $\beta = 0$ (without mutation rate correlation or substitution rate heterogeneity)

	Diversity (%)									
m	1		5		10		15		25	
5	0.4	0.4	1.6	0.9	3.6	1.7	4.2	2.4	5.1	3.7
10	0.1	0.0	3.1	1.5	4.6	3.5	3.9	3.2	4.7	4.0
15	0.2	0.0	5.5	3.8	5.7	4.7	5.4	4.5	4.0	3.8
25	0.3	0.2	4.6	2.9	4.8	4.3	4.5	3.8	4.5	4.1
50	0.8	0.1	5.9	4.5	4.1	3.8	5.7	5.6	5.7	5.3

The columns for each parameter pair represent $P_N(\Phi_w)$ and $P_P(\Phi_w)$, respectively.

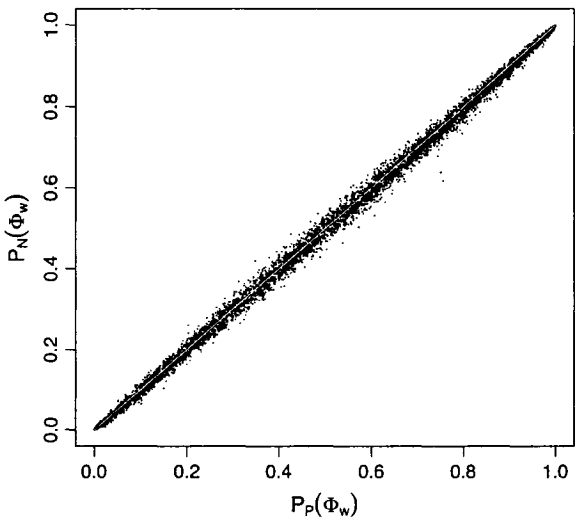


FIGURE 3.—Comparison of *P*-values obtained using the permutation test (horizontal axis) to analytical *P*-values (vertical axis) when $\rho = 0$ and $\beta = 0$. Points with <15 samples and <10% sequence divergence are not shown (see Table 2).

to viral genomic samples, were analyzed on a Mac G4 desktop computer. The time taken to analyze each alignment was ~20 sec using Φ_w without the permutation test, 30 sec using Φ_w with the permutation test, 7 min with the linkage disequilibrium methods (using LDHat), and 8 hr using the likelihood permutation test of LDHat (using a precomputed likelihood file). For longer alignments, however, the permutation test becomes impractical even for Φ_w and in these cases analytical *P*-values are the only way to practically test for recombination. It is worth noting that since the power to detect recombination increases as a function of sequence length (WIUF *et al.* 2001), this constitutes an important advantage for the Φ_w-test, since faint recombinant signals may be detectable using only very long sequences.

Power: Figure 4 shows the power to detect recombination for Φ_w, Max χ^2, NSS, the LPT in LDHat, and two measures of correlation of linkage disequilibrium with distance (r^2 and $|D'|$), when the rate of recombination ρ is greater than zero, for two different sample sizes ($m =$ 10 and $m = 50$). Two principal types of genealogies were created: with and without population growth. If there is population growth, the genealogies created will be more star-like with long branches at the leaves (GRIFFITHS and TAVARÉ 1998; WIUF *et al.* 2001). If there is no population growth, there are short branches at the tip but long branches at the root. When genealogies are more star-like, recurrent mutations will tend to mask the initial recombination, and the recombination events are best considered to be "ancestral."

The top rows of Figure 4, a and b, show that without population growth ($\beta = 0$), all six methods performed similarly, although overall Φ_w is the most powerful method with a large number of samples. Without population growth, the power to detect recombination of all six methods generally increases as a function of both sequence diversity and the rate of recombination, similar to earlier observations (POSADA and CRANDALL 2001; WIUF *et al.* 2001). A notable exception is the LPT for which there is a slight decline in power when the mean sequence diversity reaches 10%. At this point, a likelihood file with a value of θ per site of 0.1 was used rather than a likelihood file with a value of θ per site of 0.001. However, when the sequence diversity reaches 10%, the expected value of θ per site is ~0.12, suggesting that a value of θ per site of 0.1 is a better choice. Nonetheless, more power may be obtained by using a gross underestimate of θ, although previous work has demonstrated a relative insensitivity of the LPT to a specific estimate of θ (MCVEAN *et al.* 2002).

The top rows of Figure 4, a and b, suggest that the Φ_w method performs similarly to the linkage disequilibrium approaches when there is very little sequence diversity (*e.g.*, $p = 1\%$), despite the fact that the test is too conservative in these circumstances (Table 2). For very little sequence diversity (*i.e.*, $p = 1\%$), the coalescent-based method LPT is the most powerful method in constant-size populations, but has about the same power as Φ_w for growing populations. However, the results suggest that all methods may underestimate the presence of recombination if few sequences are present with very little divergence, especially in an expanding population (or "star-like" genealogy).

By comparing the bottom rows of Figure 4, a and b, to the top rows of Figure 4, a and b, it is evident that detecting the presence of recombination under population growth ($\beta = 5000$) is a more difficult task than detecting the presence of recombination without population growth ($\beta = 0$). Of all six methods, the bottom rows of Figure 4, a and b, suggest that Φ_w is much better at detecting recombination under population growth than Max χ^2, NSS, the coalescent-based LPT, or the linkage disequilibrium approaches. For the coalescent-based LPT, it is worth noting that population growth could be incorporated in the method in the future, possibly increasing power. The decline of linkage disequilibrium in expanding populations using r^2 is consistent with previous observations (SLATKIN 1994; MCVEAN 2002), but the results suggest that the performance of the $|D'|$ statistic is similar. The results for the Φ_w-test suggest that subsequent mutations do not "mask" the recombinant signal for this method. Interestingly, this is similar behavior to the RECPARS method (HEIN 1993; WIUF *et al.* 2001) and may be of particular importance when trying to determine ancestral recombination between diverged genotypes. The results also suggest that the Φ_w-statistic can be used to distinguish between star-like genealogies due to population growth and star-like genealogies due to recombination (SCHIERUP and HEIN 2000b).

A comparison of the top row of Figure 4a to the top row of Figure 4b reveals that an increase in sample size

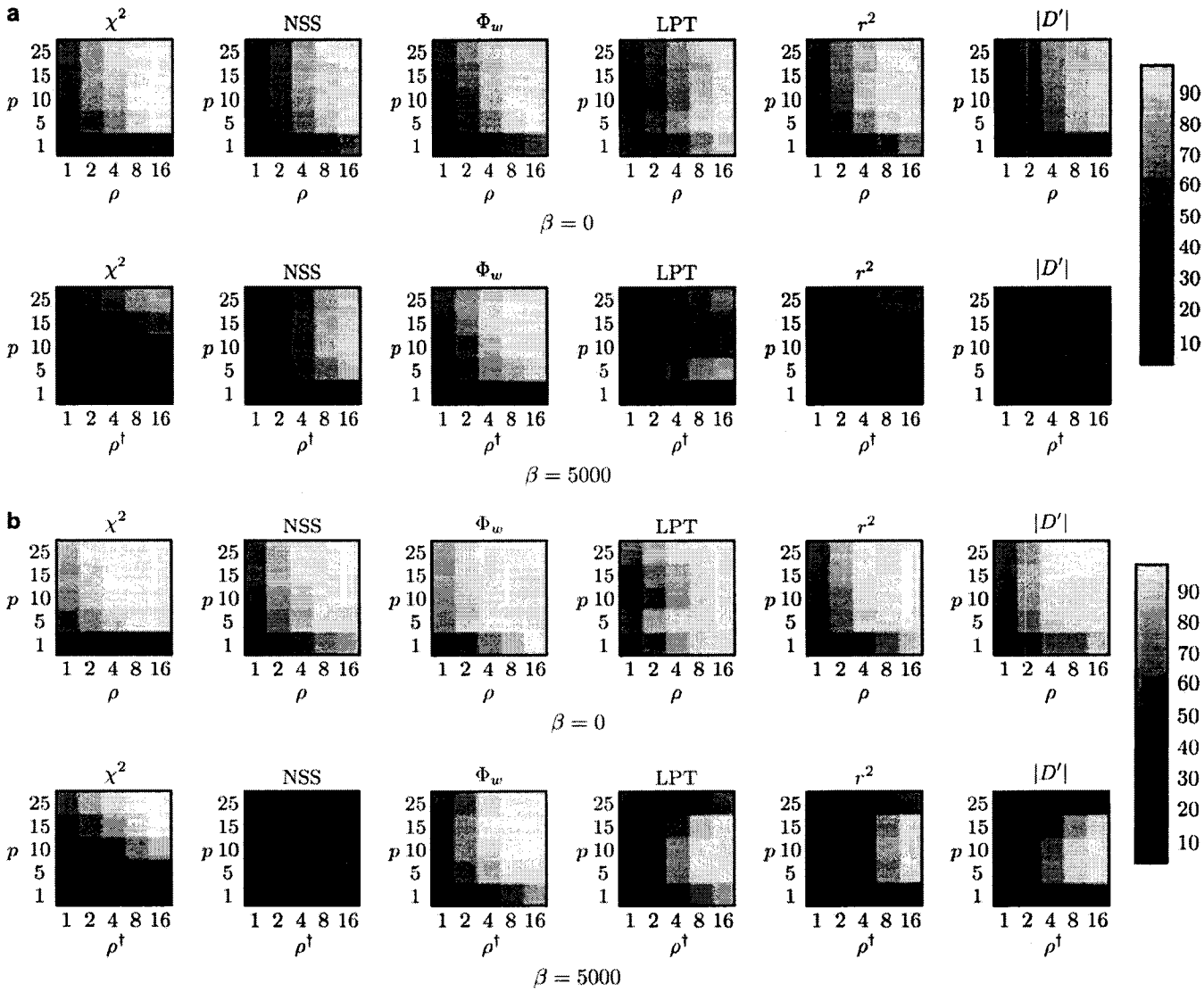


FIGURE 4.—Power to detect recombination for (a) $m = 10$ and (b) $m = 50$ samples for six different methods with (a and b, bottom rows) and without (a and b, top rows) population growth. The horizontal axis varies the rate of recombination whereas the vertical axis varies the amount of sequence diversity. Each cell represents the outcome of 1000 replicates with cells with lighter shading indicating increased power. The value $\rho^\dagger$ refers to the value of ρ used to give the same expected number of recombinations under population growth.

from $m = 10$ to $m = 50$ causes an increase in the ability of all six methods to infer recombination when there is no population growth ($\beta = 0$). For population growth (the bottom rows of Figure 4, a and b), the power to detect recombination for the NSS statistic for actually decreases sharply from $m = 10$ to $m = 50$. But for the other five tests, the power to detect recombination generally increases when moving from $m = 10$ to $m = 50$ even under population growth. These results expand upon some previous observations (WIUF *et al.* 2001).

Under a neutral coalescent model with recombination, it is possible to use a likelihood-ratio test to determine whether the hypothesis of no recombination ($\rho = 0$) should be rejected at a given significance level (KUHNER *et al.* 2000; BROWN *et al.* 2001). However, even when data are simulated according to the neutral coalescent with low levels of recombination, the hypothesis $\rho = 0$ is rejected only a limited proportion of the time (BROWN *et al.* 2001). However, such a simulation represents an ideal situation, where the likelihood-ratio test is guaranteed to be the most powerful (BROWN *et al.* 2001) and the model used to infer ρ is identical to the model used to generate samples. This suggests that it might be difficult for any test to correctly infer the presence of recombination for very low recombination rates. Additionally, a theoretical analysis shows that generating small sets of samples using a low rate of recombination produces only a limited number of incompatibilities (WIUF *et al.* 2001). It is thus possible that full-likelihood approaches (KUHNER *et al.* 2000; FEARNHEAD and DONNELLY 2001) or a phylogenetic network (HUSON and BRYANT 2006) approach could be particularly useful

TABLE 3

Power to detect recombination using Φ_w with a high rate of recombination $\rho = 128$

Diversity (%)	No. of samples $m = 10$ (%)	$m = 50$ (%)
1	68	99
5	100	100
10	100	100
15	100	100
25	100	100

to determine whether there is any possibility of recombination when only a weak recombinant signal exists.

Table 3 demonstrates that Φ_w can detect recombination even under extremely high recombination rates ($\rho = 128$). Except for low sequence diversity ($p = 1\%$), the presence of recombination is correctly inferred each time. But even for low sequence diversity, the presence of recombination can be inferred nearly every time by increasing the sample size from $m = 10$ to $m = 50$.

It is worth noting that the Φ_w-statistic can also be calculated without the refined incompatibility score, but using only the traditional notion of compatibility. For cases without population growth ($\beta = 0$), the results are almost identical (results not shown). On the other hand, with population growth ($\beta = 5000$), there is an increase in power using the refined incompatibility score when the number of samples is large (*e.g.*, $m = 50$) and there is some recurrent mutation. For a rate of recombination of $\rho = 1$, a sample size of 50, and exponential growth, the gains in power using the refined incompatibility score rather than the compatibility score were 2, 5, and 12% for mean pairwise sequence divergences of 10, 15, and 25%, respectively. Similar results are obtained for $\rho = 2$ but not for higher rates of recombination (results not shown). This suggests that the refined incompatibility score is a useful extension to the traditional notion of compatibility especially for large sample sizes with sites that experience recurrent mutations.

For no population growth, the Φ_w-test and the linkage disequilibrium approaches perform similarly, although Φ_w is more powerful for a large number of samples. However, Φ_w is applicable even if the samples are from different species or different populations, whereas the linkage disequilibrium and coalescent approaches are not (Tsaousis *et al.* 2005). Under population growth, however ($\beta = 5000$), only Φ_w continues to consistently infer the presence of recombination as the power of the other five methods suffers sharp declines. This suggests that, of all six methods, Φ_w has the greatest flexibility in detecting recombination in the different circumstances studied.

False positives: Of particular concern for any test for recombination is the effect of confounding processes such as substitution rate heterogeneity and autocorrelated substitution rates. Autocorrelation of substitution rates implies that the rate of substitution of one site is not independent of the rate of substitution of a neighboring site and can create "mutational hot spots" within a sequence. This can potentially create the same patterns as recombination.

Figure 5 shows the proportion of false positives for Max χ^2 and NSS when there is no recombination ($\rho = 0$) but "mosaic" sequences are artificially induced by using a range of autocorrelated substitution rates. Figure 5 shows that both Max χ^2 and NSS falsely infer the presence of recombination >50% of the time in certain cases. The results for the linkage disequilibrium, likelihood permutation test, and Φ_w are omitted from Figure 5 since these methods did not falsely infer

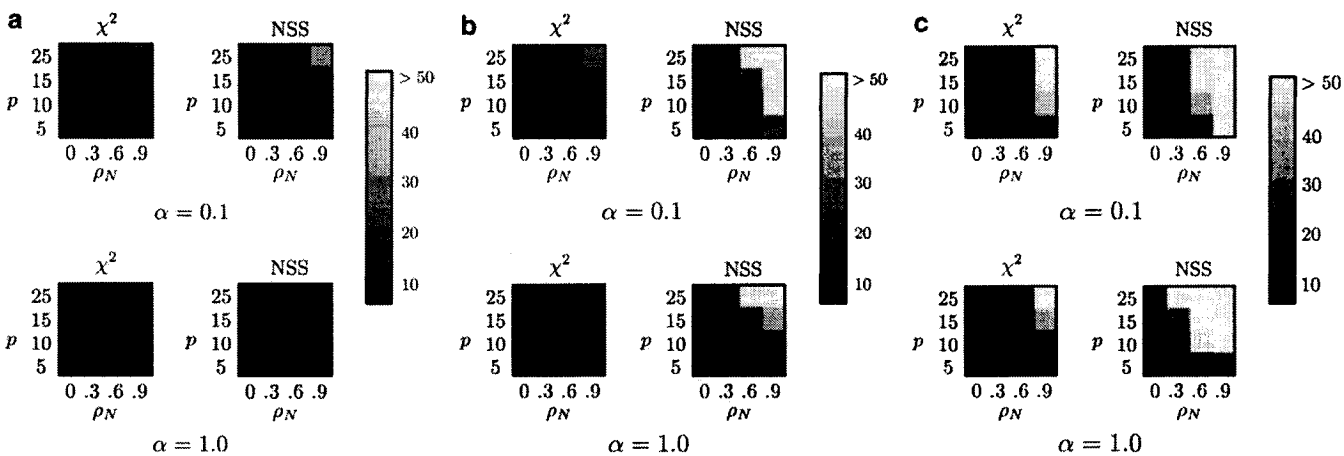


FIGURE 5.—Percentage of false positives for (a) $m = 10$ samples (with $\beta = 5000$), (b) $m = 50$ samples (with $\beta = 0$), and (c) $m = 50$ samples (with $\beta = 5000$), for Max χ^2 and NSS, with extreme rate heterogeneity (top row) and moderate rate heterogeneity (bottom row). The horizontal axis varies the substitution rate correlation whereas the vertical axis varies the amount of sequence diversity. Each cell represents the outcome of 1000 replicates with cells with lighter shading indicating a higher percentage of false positives. The results for Φ_s, r^2, and $|D'|$ are omitted since these approaches did not falsely infer recombination >7% of the time for any of the conditions, but Table 4 shows a number of these results for Φ_w.

TABLE 4

Proportion of times recombination is falsely inferred using Φ_w with substitution rate heterogeneity $\alpha = 0.1$, mutation rate correlation, and sample size $m = 50$

	Mutation rate correlation							
Diversity (%)	0		0.3		0.6		0.9	
1	2.0	3.6	2.5	3.6	2.6	3.9	1.1	3.8
5	4.9	4.7	5.8	4.5	4.7	3.3	3.0	1.0
10	4.1	5.6	4.7	4.6	4.8	3.0	1.8	1.5
15	4.9	4.0	4.5	4.7	3.8	4.5	2.9	1.8
25	5.3	4.0	3.7	3.5	4.1	3.9	3.4	2.1

The columns for each parameter pair represent the outcomes for $\beta = 0$ and $\beta = 5000$, respectively.

recombination >7% of the time, although Table 4 shows this information for Φ_w. Table 4 shows that the Φ_w-statistic did not infer recombination >6% of the time when recombination was falsely inferred >50% of the time using both Max χ^2 and NSS. Although the global model of substitution rate autocorrelation employed by this study is quite simple since it ignores codon positions and substitution rate correlation within local patterns of substitution (McVean 2001), it nonetheless provides a guide to the effect of autocorrelated substitution rates.

The problem of false positives in NSS and Max χ^2 is most severe for large sample sizes (*e.g.*, $m = 50$), both under constant-size populations (Figure 5b) and under population growth (Figure 5c). Although the problem is in general greater for higher substitution heterogeneity (Figure 5, top rows) it is also a problem with lower substitution rate heterogeneity (Figure 5, bottom rows).

The level of false positives of both NSS and Max χ^2 suggests caution in interpreting evidence for recombination, especially when autocorrelated rates are an issue. For instance, inferring the presence of recombination in mitochondrial DNA should be done cautiously as substitution rate correlation is known (Yang 1995; Nielsen 1997).

The results using Φ_w contrast strongly with the results using the NSS (which is also compatibility based). This is likely due to the difference in the statistics themselves. The Φ_w-statistic uses compatibility between closely linked sites directly whereas the NSS statistic measures clustering within a compatibility matrix. As the clustering can be caused by substitution rate correlation, and not only by recombination, this might explain the difference between the two statistics. For Max χ^2 the problem is possibly due to pairs of sequences that differ greatly on one side of a site (due to high mutation) but share a great degree of similarity on the other side of a site (due to low mutation). Local "bursts" of mutation (McVean 2001) likely exacerbate the problem, especially for linkage disequilibrium approaches that are based on allele frequencies at different sites.

Empirical data: The general information concerning the empirical data sets is summarized in Table 1. Tables 5 and 6 show the results of tests for recombination on all the empirical data sets. In addition to the results obtained using the Φ_w-statistic, results using Max χ^2 (Maynard Smith 1992), NSS (Jakobsen and Easteal 1996), correlation of r^2 and $|D'|$ with distance (Lewontin 1964; Hill and Robertson 1968), and a LPT (McVean *et al.* 2002) are shown. The estimates of ρ for the population level data sets were obtained using LDHat (McVean *et al.* 2002). Tests for recombination within populations (*i.e.*, r^2, $|D'|$, and LPT) were not applied to data sets that contained individuals from different species.

Recombinant examples: Table 5 shows that the null hypothesis of no recombination is rejected by all tests for most of the suspected recombinant data sets, including the Candida example that had very little sequence diversity (0.7%). Whereas a lack of sequence diversity in the simulations made recombination harder to detect, this may be partially overcome by using longer alignments, such as that for the Candida example, which had 2553 sites. Interestingly, the null hypothesis of no recombination was not universally rejected for two of the bacterial data sets: Cowdria and *Helicobacter pylori*. For

TABLE 5

Analysis of suspected recombinant data sets

Data set	ρ^a	$\Phi_w{}^{b,c}$	χ^2	NSS	$r^{2a,d}$	$\lvert D'\rvert^{a,d}$	LPT^{a,d,e}
Candida	16	2.4×10^{-15}* (0.000*)	0.000*	0.000*	0.000* (0.000*)	0.122 (0.001)	0.000* (0.000*)
Rana	—	5.5×10^{-31}* (0.000*)	0.000*	0.000*	—	—	
Cowdria	17	3.8×10^{-5}* (0.000*)	0.041*	0.001*	0.167 (0.039*)	0.043* (0.029*)	0.000* (0.001*)
H. pylori	≥100	9.3×10^{-3}* (0.004*)	0.158	0.330	0.125 (0.000*)	0.536 (0.003*)	0.000* (0.000*)

* $P < 0.05$.

[a] Calculated on sites with only two alleles segregating with LDHat.

[b] Each pair shows *P*-values calculated analytically and using a permutation test, respectively.

[c] w was set to 100 for all tests.

[d] Terms in parentheses show results on sites with minor allele frequencies >0.1.

[e] Denotes the value of a likelihood permutation test calculated in LDHat.

TABLE 6

Analysis of possibly recombinant data sets

Data set	ρ^a	$\Phi_w^{b,c}$	χ^2	NSS	$r^{2a,d}$	$\lvert D' \rvert^{a,d}$	LPT^{a,d,e}
Norovirus	23 (21)	0.002* (0.003*)	0.025*	0.237	0.029* (0.574)	0.868 (0.340)	0.022* (0.026*)
Apodemus	—	0.135 (0.151)	0.274	0.006*	—	—	—
Boletales	—	0.934 (0.931)	0.003*	0.000*	—	—	—
Wolbachia	0 (2)	0.086 (0.103)	0.566	0.108	0.049* (0.019*)	0.286 (0.204)	0.709 (0.090)

* $P < 0.05$.

[a] Calculated on sites with only two alleles segregating.

[b] Each pair shows *P*-values calculated analytically and using a permutation test, respectively.

[c] w was set to 100 for all tests.

[d] Terms in parentheses show results on sites with minor allele frequencies >0.1.

[e] Denotes the value of a likelihood permutation test calculated in LDHat.

these two bacterial examples, evidence for recombination was found using the Φ_w-statistic as well as the coalescent-based likelihood permutation test. However, recombination was detected in the Cowdria example using the correlation of distance with r^2 only after sites with minor alleles were removed. Moreover, in the *H. pylori* data set neither NSS nor Max χ^2 found significant evidence for recombination. This could be due to the high suspected rate of recombination in the *H. pylori* example, which has conditions approaching linkage equilibrium (SUERBAUM *et al.* 1998). The linkage disequilibrium methods seem to be highly sensitive to sites with low allele frequencies and consistent results are obtained only after the removal of these sites.

Possibly recombinant examples: The results obtained from the data sets for which the status of recombination is debated are quite interesting (Table 6). For the Norovirus example, evidence of recombination is found using Φ_w, Max χ^2, and the LPT. There is some evidence of recombination found with r^2, but after sites with minor allele frequencies <0.1 are removed no further evidence is found by the linkage disequilibrium methods. Since the samples came from a number of different cities, it could be that evidence of recent recombination is weakened by removing these sites. However, the LPT finds evidence of recombination regardless of whether or not these sites are removed.

For the bacterial symbiont nematode Wolbachia, there is little prior reason to suspect recombination (JIGGINS 2002). Nonetheless, evidence for recombination is found using correlation of r^2 with distance and marginal evidence for recombination is found by using the likelihood permutation test when sites with minor alleles frequencies <0.1 are removed. The results obtained using the Φ_w-statistic also suggest that there is marginal evidence for recombination with Wolbachia. The possible presence of recombination in Wolbachia should be tested further using more data.

Recombination in the animal mitochondrial DNA of Apodemus was first proposed (LADOUKAKIS and ZOUROS 2001) and then disputed (MAYNARD SMITH and SMITH 2002). Tests for recombination using Φ_w and Max χ^2 indicate that there is little evidence for recombination, although the NSS statistic does find evidence for recombination. The evidence for recombination within Apodemus using the Max χ^2-test is even weaker here than in previous studies (MAYNARD SMITH and SMITH 2002), possibly due to the fact that this implementation of the Max χ^2-test uses a "fixed window size." Given the high level of false positives of NSS, the results suggest that evidence for recombination within Apodemus is lacking.

For the fungal Boletales, results using the Φ_w-statistic are quite distinct from the results obtained using both the NSS and the Max χ^2-statistic. The Φ_w-based tests find no evidence for recombination whereas both other tests find strong evidence for recombination. Interestingly, although most other methods for detecting recombination find evidence for recombination within this data set, Geneconv (SAWYER 1989), another powerful sequence-based test for recombination, does not (POSADA 2002).

One possibility for the Boletales data set is that the Φ_w-statistic is too conservative and produced a type II error ("false negative"). The Boletales data set is a saturated data set with a strong A + T bias (KRETZER and BRUNS 1999). The strong A + T bias results in an estimated transition/transversion ratio of 0.4. Simulations show, however, that even under such conditions, there is reason to believe that recombination will still create distinct patterns of compatibility and incompatibility that should be detectable using the Φ_w-statistic (results not shown). Moreover, simulations indicate that the Φ_w-statistic appears to be more powerful than the NSS statistic (which is also compatibility based), suggesting that a type II error for the Φ_w-statistic, but not for the NSS statistic, is unlikely.

Another possibility for the Boletales example is that both Max χ^2 and the NSS statistic are producing type I errors, which, according to the simulations, autocorrelated substitution rates might induce. To test this, a parametric bootstrap with 1000 replicates simulating codons (with no recombination) was performed using a substitution rate heterogeneity of 1.31 and global

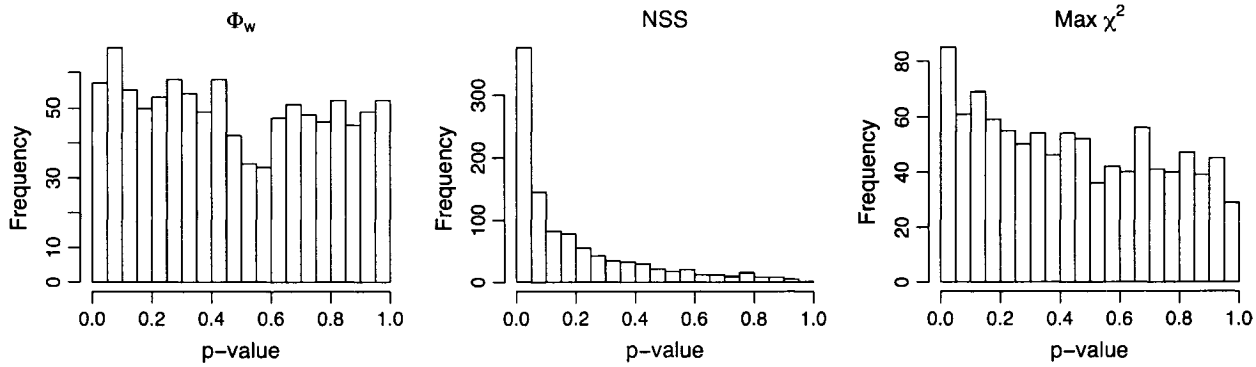


FIGURE 6.—Distribution of *P*-values inferred by the Φ_w-statistic, the NSS statistic, and the Max χ^2-statistic. The results are obtained on the basis of 1000 parametric bootstraps under conditions observed for the Boletales example. None of the replicates contained recombination but the substitution rate autocorrelation was set to $\rho_N = 0.35$ and substitution rate heterogeneity was set to $\alpha = 1.31$.

substitution rate correlation $\rho_G = 0.35$ as estimated from the data set. Figure 6 shows the distribution of estimated *P*-values obtained on the 1000 replicates using the Max χ^2-statistic, NSS statistic, and the Φ_w-statistic. Recombination was inferred 5.7% of the time using the Φ_w-statistic, 8.5% of the time with the Max χ^2-statistic, and 37.5% of the time using the NSS statistic. Since none of the replicates contained recombination, the *P*-values for each of the three methods should follow a uniform distribution. Figure 6 shows that the parametric bootstrap creates conditions similar to recombination for both Max χ^2 and NSS [a one-sided Kolmogorov–Smirnov test (MASSEY 1951) rejects the uniform distribution at a significance level of 10^{-7} for both Max χ^2 and NSS but fails to find any evidence to reject the uniform distribution for Φ_w]. Whereas the results for Max χ^2 are less striking than those for NSS, the parametric bootstrap fails to account for local patterns of mutation (HEY 2000; MCVEAN 2001; MCVEAN *et al.* 2002), which are likely to exacerbate the observed bias. These results suggest that there is reason to doubt the validity of the inferences of Max χ^2 and NSS concerning the presence of recombination in the Boletales data set.

Conclusion: We have presented a simple, powerful test for detecting recombination that can be used regardless of sample history. The approach is very general (*e.g.*, does not assume a single population) and aims to determine simply whether there is a recombinant signal present within the sequences. In contrast to two other general tests, Max χ^2 and NSS, our test does not falsely infer the presence of recombination because of mutation rate correlation (which is present in some mitochondrial DNA). Interestingly, our approach performs very well even in the presence of population growth, in contrast to methods based on linkage disequilibrium (r^2 and $|D'|$), a coalescent-based likelihood permutation test (from LDHat), Max χ^2, and NSS. Our method can be used by itself, or to validate the visual presence of recombination from a phylogenetic network approach, or to independently verify the presence of recombination if a positive estimate of the rate of recombination is obtained. The approach may be particularly useful in distinguishing recurrent mutation from recombination when assumptions such as a single, randomly mating, and constant-size population are not met. The test can be used easily when many sequences and sites are present because of its computational efficiency and indeed is more powerful in such circumstances. A program implementing our test as well as both Max χ^2 and NSS is available as a stand-alone program at the following address: http://www.mcb.mcgill.ca/~trevor. The test is also implemented in SplitsTree 4.2, available at http://www.splitstree.org.

T.B. thanks Kirk and Rachel Bevan, Scott Bunnell, Daniel Huson, and Russell Steele, as well as the two anonymous referees for a number of helpful suggestions that greatly improved the manuscript. T.B. is supported by the National Science Engineering and Research Council (NSERC) (postgraduate scholarship B) and by Le Fonds québécois de la recherche sur la nature et les technologies (FQRNT grant 2003-NC-81840). D.B. is supported in part by NSERC (grant 238975-01). H.P. acknowledges Génome Québec.

LITERATURE CITED

ANDERSON, J. B., C. WICKENS, M. KHAN, L. E. COWEN, N. FEDERSPIEL *et al.*, 2001 Infrequent genetic exchange and recombination in the mitochondrial genome of *Candida albicans*. J. Bacteriol. **183**(3): 865–872.

AWADALLA, P., 2003 The evolutionary genomics of pathogen recombination. Nat. Rev. Genet. **4**(1): 50–60.

AWADALLA, P., A. EYRE-WALKER and J. M. SMITH, 1999 Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science **286**(5449): 2524–2525.

BROWN, C. J., E. C. GARNER, A. KEITH DUNKER and P. JOYCE, 2001 The power to detect recombination using the coalescent. Mol. Biol. Evol. **18**(7): 1421–1424.

BRUEN, T., and D. BRYANT, 2006 A subdivision approach to maximum parsimony. Ann. Combinator. (in press).

CAMIN, J. H., and R. R. SOKAL, 1965 A method for deducing branching sequences in phylogeny. Evolution **19**(3): 311–326.

CASELLA, G., and R. L. BERGER, 2001 *Statistical Inference*. Duxbury Press, Belmont, CA.

CRANDALL, K. A., and A. R. TEMPLETON, 1999 Statistical approaches to detecting recombination, pp. 153–176 in *The Evolution of HIV*, edited by K. A. CRANDALL. Johns Hopkins University Press, Baltimore.

DROUIN, G., F. PRAT, M. ELL and G. D. CLARKE, 1999 Detecting and characterizing gene conversions between multigene family members. Mol. Biol. Evol. **16**(10): 1369–1390.

FEARNHEAD, P., and P. DONNELLY, 2001 Estimating recombination rates from population genetic data. Genetics **159:** 1299–1318.

FELSENSTEIN, J., 2004 *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

GRASSLY, N. C., and E. C. HOLMES, 1997 A likelihood method for the detection of selection and recombination using nucleotide sequences. Mol. Biol. Evol. **14**(3): 239–247.

GRASSLY, N. C., P. H. HARVEY and E. C. HOLMES, 1999 Population dynamics of HIV-1 inferred from gene sequences. Genetics **151:** 427–438.

GRIFFITHS, R. C., and P. MARJORAM, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. **3**(4): 479–502.

GRIFFITHS, R. C., and S. TAVARÉ, 1998 The age of a mutation in a general coalescent tree. Stoch. Models **14:** 273–295.

HAGENBLAD, J., and M. NORDBORG, 2002 Sequence variation and haplotype structure surrounding the flowering time locus FRI in *Arabidopsis thaliana*. Genetics **161:** 289–298.

HAYDON, D. T., A. D. S. BASTOS and P. AWADALLA, 2004 Low linkage disequilibrium indicative of recombination in foot-and-mouth disease virus gene sequence alignments. J. Gen. Virol. **85:** 1095–1100.

HEIN, J., 1990 Reconstructing evolution of sequences subject to recombination using parsimony. Math. Biosci. **98**(2): 185–200.

HEIN, J., 1993 A heuristic method to reconstruct the history of sequences subject to recombination. J. Mol. Evol. **36**(4): 396–405.

HEIN, J., M. H. SCHIERUP and C. WIUF, 2005 *Gene Genealogies, Variation and Evolution*. Oxford University Press, London/New York/Oxford.

HEY, J., 2000 Human mitochondrial DNA recombination: Can it be true? Trends Ecol. Evol. **15**(5): 181–182.

HEY, J., and J. WAKELEY, 1997 A coalescent estimator of the population recombination rate. Genetics **145:** 833–846.

HILL, W., and A. ROBERTSON, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **33:** 54–78.

HUDSON, R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

HUDSON, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159:** 1805–1817.

HUDSON, R. R., and N. L. KAPLAN, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

HUSON, D. H., and D. BRYANT, 2006 Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. **23:** 254–267.

INNAN, H., and M. NORDBORG, 2002 Recombination or mutational hot spots in human mtDNA? Mol. Biol. Evol. **19**(7): 1122–1127.

JAKOBSEN, I. B., and S. EASTEAL, 1996 A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. Comput. Appl. Biosci. **12**(4): 291–295.

JIGGINS, F. M., 2002 The rate of recombination in Wolbachia bacteria. Mol. Biol. Evol. **19**(9): 1640–1643.

JUKES, T. H., and C. R. CANTOR, 1969 *Mammalian Protein Metabolism*, Vol. III, pp. 21–132. Academic Press, New York/London.

KIMURA, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61:** 893–903.

KINGMAN, J., 1982 The coalescent. Stoch. Proc. Appl. **13:** 235–248.

KRETZER, A. M., and T. D. BRUNS, 1999 Use of atp6 in fungal phylogenetics: an example from the boletales. Mol. Phylogenet. Evol. **13**(3): 483–492.

KUHNER, M. K., J. YAMATO and J. FELSENSTEIN, 2000 Maximum likelihood estimation of recombination rates from population data. Genetics **156:** 1393–1401.

LADOUKAKIS, E. D., and E. ZOUROS, 2001 Recombination in animal mitochondrial DNA: evidence from published sequences. Mol. Biol. Evol. **18**(11): 2127–2131.

LE QUESNE, W. J., 1969 A method of selection of characters in numerical taxonomy. Syst. Zool. **18**(2): 201–205.

LEWONTIN, R., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49:** 49–67.

MARTIN, D., and E. RYBICKI, 2000 RDP: detection of recombination amongst aligned sequences. Bioinformatics **16**(6): 562–563.

MARTIN, Y., G. GERLACH, C. SCHLOTTERER and A. MEYER, 2000 Molecular phylogeny of European muroid rodents based on complete cytochrome b sequences. Mol. Phylogenet. Evol. **16**(1): 37–47.

MASSEY, F. J., 1951 The Kolmogorov-Smirnov test for goodness of fit. J. Am. Stat. Assoc. **46**(253): 68–78.

MAYNARD SMITH, J., 1992 Analyzing the mosaic structure of genes. J. Mol. Evol. **34**(2): 126–129.

MAYNARD SMITH, J., and N. H. SMITH, 2002 Recombination in animal mitochondrial DNA. Mol. Biol. Evol. **19**(12): 2330–2332.

MCGUIRE, G., and F. WRIGHT, 2000 TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. Bioinformatics **16:** 130–134.

MCVEAN, G., P. AWADALLA and P. FEARNHEAD, 2002 A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics **160:** 1231–1241.

MCVEAN, G. A., 2001 What do patterns of genetic variability reveal about mitochondrial recombination? Heredity **87:** 613–620.

MCVEAN, G. A. T., 2002 A genealogical interpretation of linkage disequilibrium. Genetics **162:** 987–991.

MININ, V. N., K. S. DORMAN, F. FANG and M. A. SUCHARD, 2005 Dual multiple change-point model leads to more accurate recombination detection. Bioinformatics **21:** 3034–3042.

MIYASHITA, N., and C. H. LANGLEY, 1988 Molecular and phenotypic variation of the white locus region in *Drosophila melanogaster*. Genetics **120:** 199–212.

MYERS, S. R., and R. C. GRIFFITHS, 2003 Bounds on the minimum number of recombination events in a sample history. Genetics **163:** 375–394.

NIELSEN, R., 1997 Site-by-site estimation of the rate of substitution and the correlation of rates in mitochondrial DNA. Syst. Biol. **46**(2): 346–353.

NIELSEN, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154:** 931–942.

PENNY, D., and M. HENDY, 1986 Estimating the reliability of evolutionary trees. Mol. Biol. Evol. **3**(5): 403–417.

PIGANEAU, G., M. GARDNER and A. EYRE-WALKER, 2004 A broad survey of recombination in animal mitochondria. Mol. Biol. Evol. **21**(12): 2319–2325.

POSADA, D., 2001 Unveiling the molecular clock in the presence of recombination. Mol. Biol. Evol. **18**(10): 1976–1978.

POSADA, D., 2002 Evaluation of methods for detecting recombination from DNA sequences: empirical data. Mol. Biol. Evol. **19**(5): 708–717.

POSADA, D., and K. A. CRANDALL, 2001 Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc. Natl. Acad. Sci. USA **98**(24): 13757–13762.

POSADA, D., and K. A. CRANDALL, 2002 The effect of recombination on the accuracy of phylogeny estimation. J. Mol. Evol. **54**(3): 396–402.

ROHAYEM, J., J. MUNCH and A. RETHWILM, 2005 Evidence of recombination in the norovirus capsid gene. J. Virol. **79**(8): 4977–4990.

SAITOU, N., and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**(4): 406–425.

SAWYER, S., 1989 Statistical tests for detecting gene conversion. Mol. Biol. Evol. **6**(5): 526–538.

SCHAEFFER, S. W., and E. L. MILLER, 1993 Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. Genetics **135:** 541–552.

SCHIERUP, M. H., and J. HEIN, 2000a Consequences of recombination on traditional phylogenetic analysis. Genetics **156:** 879–891.

SCHIERUP, M. H., and J. HEIN, 2000b Recombination and the molecular clock. Mol. Biol. Evol. **17**(10): 1578–1579.

SLATKIN, M., 1994 Linkage disequilibrium in growing and stable populations. Genetics **137:** 331–336.

SLATKIN, M., and R. R. HUDSON, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555–562.

SNEATH, P., M. SACKIN and R. AMBLER, 1975 Detecting evolutionary incompatibilities from protein sequences. Syst. Zool. **24**(3): 311–332.

SONG, Y. S., and J. HEIN, 1999 On the minimum number of recombination events in the evolutionary history of DNA sequences. J. Math. Biol. **48**(2): 160–186.

SUERBAUM, S., J. M. SMITH, K. BAPUMIA, G. MORELLI, N. H. SMITH *et al.*, 1998 Free recombination within *Helicobacter pylori.* Proc. Natl. Acad. Sci. USA **95**(21): 12619–12624.

SUMIDA, M., M. OGATA and M. NISHIOKA, 2000 Molecular phylogenetic relationships of pond frogs distributed in the Palearctic region inferred from DNA sequences of mitochondrial 12S ribosomal RNA and cytochrome b genes. Mol. Phylogenet. Evol. **16**(2): 278–285.

SWOFFORD, D. L., 1998 *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods).* Sinauer Associates, Sunderland, MA.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

TSAOUSIS, A. D., D. P. MARTIN, E. D. LADOUKAKIS, D. POSADA and E. ZOUROS, 2005 Widespread recombination in published animal mtDNA sequences. Mol. Biol. Evol. **22**(4): 925–933.

UZZELL, T., and K. W. CORBIN, 1971 Fitting discrete probability distributions to evolutionary events. Science **172:** 1089–1096.

WALL, J. D., 2000 A comparison of estimators of the population recombination rate. Mol. Biol. Evol. **17**(1): 156–163.

WEILLER, G. F., 1998 Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. Mol. Biol. Evol. **15**(3): 326–335.

WEIR, B., and W. HILL, 1986 Nonuniform recombination within the human beta-globin gene cluster. Am. J. Hum. Genet. **38**(5): 776–781.

WIUF, C., and J. HEIN, 2000 The coalescent with gene conversion. Genetics **155:** 451–462.

WIUF, C., T. CHRISTENSEN and J. HEIN, 2001 A simulation study of the reliability of recombination detection methods. Mol. Biol. Evol. **18**(10): 1929–1939.

YANG, Z., 1993 Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10**(6): 1396–1401.

YANG, Z., 1995 A space-time process model for the evolution of DNA sequences. Genetics **139:** 993–1005.

YANG, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**(5): 555–556.

Communicating editor: M. VEUILLE

APPENDIX A

The normal approximation to the permutation test requires calculation of the expectation and variance of the Φ_{w} statistic under permutations of the alignment. This section contains derivations for both the mean and the variance and outlines how to compute both values efficiently. Again, assume that the proportion of informative sites is q and let w be a fixed width (in bases). Throughout this section, let $k = wq$.

Let $M = (M_{i,j})$ be a given $n \times n$ refined incompatibility matrix. Note that M is symmetric. Let $I = \{1, \ldots, n\}$ be an index set. Let σ be any permutation of the index set, and define a permutation of the matrix as $\sigma(M) = (M_{\sigma(i),\sigma(j)})$.

Define the sample space Ω by $\Omega = \{\sigma(M): \sigma \in S_n\}$. Assume that every permutation σ is equally likely. Define an $n \times n$ random matrix $X:\Omega \to \mathbb{R}^{n\times n}$ by $X = \sigma(M)$. Note that X is symmetric, a fact that is used throughout without further mention.

Define for all $1 \leq i \leq n$: $f_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} M_{i,j}$ and $g_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} M_{i,j}^2$.

Also define $u = \sum_{i=1}^{n} f_i$, $v = \sum_{i=1}^{n} g_i$, and $w = \sum_{i=1}^{n} (f_i)^2$.

LEMMA 1. *Let X be a random matrix. Then for any arbitrary but distinct $\{i, j, k, l\}$*

$$E[X_{i,j}] = \frac{(n-2)!}{n!}u$$

$$E[X_{i,j}^2] = \frac{(n-2)!}{n!}v$$

$$E[X_{i,j}X_{i,k}] = \frac{(n-3)!}{n!}(w-v)$$

$$E[X_{i,j}X_{k,l}] = \frac{(n-4)!}{n!}(u^2 + 2v - 4w).$$

Proof. Note that a permutation σ of I can be viewed as mapping to $I \to I$. Denote the value of $\sigma(i)$ by σ_i. The total number of permutations is then $n!$. The number of permutations that have m distinct elements fixed in some mapping is $(n-m)!$ (*e.g.*, $\sigma(a_1) = b_1, \sigma(a_2) = b_2, \ldots, \sigma(a_m) = b_m$). Since every permutation is equally likely the probability of such a permutation is

$$\frac{(n-m)!}{n!}.$$

Note that every distinct pair (i, j), $i \neq j$ can be mapped to any distinct pair (a, b), $a \neq b$, by some σ. Note also that $\Pr[X_{i,j} = M_{a,b}] = \Pr[\sigma_a = i \wedge \sigma_b = j]$. Finally, for notational convenience the summation $\sum_{a=1}^{n}$ is written as $\sum_a$. Hence,

$$\begin{aligned}
E[X_{i,j}] &= \sum_a \sum_{b \neq a} M_{a,b} \Pr[\sigma_a = i \wedge \sigma_b = j] \\
&= \sum_a \sum_{b \neq a} M_{a,b} \frac{(n-2)!}{n!} \\
&= \frac{(n-2)!}{n!} u \\
E[X_{i,j}^2] &= \sum_a \sum_{b \neq a} M_{a,b}^2 \Pr[\sigma_a = i \wedge \sigma_b = j] \\
&= \frac{(n-2)!}{n!} v \\
E[X_{i,j} X_{i,k}] &= \sum_a \sum_{b \neq a} \sum_{c \neq a,b} M_{a,b} M_{a,c} \Pr[\sigma_a = i \wedge \sigma_b = j \wedge \sigma_c = k] \\
&= \frac{(n-3)!}{n!} \sum_a ((f_a)^2 - g_a) \\
&= \frac{(n-3)!}{n!} (w - v) \\
E[X_{i,j} X_{k,l}] &= \sum_{a=1} \sum_{b \neq a} \sum_{c \neq a,b} \sum_{d \neq a,b,c} M_{a,b} M_{c,d} \Pr[\sigma_a = i \wedge \sigma_b = j \wedge \sigma_c = k \wedge \sigma_d = l] \\
&= \frac{(n-4)!}{n!} \left(\left(\sum_a f_a \right)^2 + \sum_a (2g_a - 4(f_a)^2) \right) \\
&= \frac{(n-4)!}{n!} (u^2 + 2v - 4w).
\end{aligned}$$

■

Consider the statistic Φ_w defined on a random matrix X as

$$\Phi_w = \frac{2}{k(2n-k-1)} \sum_{j=1}^{k} \sum_{i=1}^{n-j} X_{i,i+j}.$$

Define (for $1 \leq a, b \leq n$)

$$P_k = \{(a, b): a < b \leq a + k\}.$$

Note that

$$|P_k| = (n-1) + (n-2), \ldots, (n-k) = \frac{k(2n-k-1)}{2}.$$

Then

$$\Phi_w = \frac{1}{|P_k|} \sum_{(a,b) \in P_k} X_{a,b}.$$

THEOREM 1. *The expectation and variance of* Φ_w *can be written as*

$$\begin{aligned}
E[\Phi_w] &= \frac{(n-2)!}{n!}(u) \\
\mathrm{Var}[\Phi_w] &= c_1 u^2 + c_2 v + c_3 w
\end{aligned}$$

(for $n \geq 2k$), where

$$c_1 = \frac{2}{3}\frac{27kn - 18k^2 + 28k^2n - 21kn^2 - 9k + 5n - 9k^3 - 11n^2 + 6n^3 + 6k^3n - 4k^2n^2}{k(k+1-2n)^2(n-1)^2(n-2)(n-3)n^2}$$

$$c_2 = \frac{2}{3}\frac{39kn - 14k^2 + 8k^2n - 15kn^2 - 21k + 19n + 3k^3 - 21n^2 + 6n^3 - 4}{k(k+1-2n)^2n(n-1)(n-2)(n-3)}$$

$$c_3 = -\frac{4}{3}\frac{-18kn - 2k^2n + 16k^2 + 6n^2 - 10n + 2 + 15k + 3k^3}{k(k+1-2n)^2n(n-1)(n-2)(n-3)}.$$

Moreover, both $E[\Phi_w]$ and $\mathrm{Var}[\Phi_w]$ can be calculated in $O(n^2)$ time.

Proof. The expectation is straightforward:

$$E[\Phi_w] = \frac{1}{|P_k|}\sum_{(a,b)\in P_k} E[X_{a,b}] = \frac{(n-2)!}{n!}u.$$

The variance is a little more involved,

$$\mathrm{Var}[\Phi_w] = \mathrm{Var}\left[\frac{1}{|P_k|}\sum_{(a,b)\in P_k} X_{a,b}\right]$$

$$= \frac{1}{|P_k|^2}\left(\sum_{(a,b)\in P_k}\mathrm{Var}[X_{a,b}] + 2\sum_{((a,b),(c,d))\in Q_k}\mathrm{Cov}[X_{a,b}X_{c,d}]\right),$$

where

$$Q_k = \{((a,\ b),\ (c,\ d)) \in P_k \times P_k :\ (a,\ b) \prec (c,\ d)\}$$

and $\prec$ denotes standard lexicographical ordering.

Note that Q_k can be partitioned into two disjoint sets $Q_{k,0}$ and $Q_{k,1}$, where $Q_{k,m} = \{((a,b),(c,d)) \in Q_k : |\{a,b\} \cap \{c,d\}| = m\}$ [by definition Q_k does not contain pairs of the type $((a,b),(a,b))$]. One way to determine $Q_{k,1}$ is to set up a recurrence.

Note that

$$P_1 = \{(1,\ 2),\ (2,\ 3),\ \ldots,\ (n-1,\ n)\}$$

so that

$$Q_{1,1} = \{((a,\ a+1),\ (a+1,\ a+2)):\ 1 \leq a \leq n-2\}.$$

Hence $|Q_{1,1}| = (n-2)$.

Next let $((a_1, a_2), (a_3, a_4)) \in Q_k - Q_{k-1}$. Then at least one $(a_1, a_2) = (a, a+k)$ or $(a_3, a_4) = (a, a+k)$ must be true. Consider the four subcases:

Case 1: $((a,b),(a,a+k))$, where $1 \leq a \leq n-k$ and $a < b < a+k$. There are precisely $(n-k)(k-1)$ terms of this type.

Case 2: $((a,a+k),(b,a+k))$, where $1 \leq a \leq n-k$ and $a < b < a+k$. Again, there are precisely $(n-k)(k-1)$ terms of this type.

Case 3: $((a,a+k),(a+k,b))$, where $1 \leq a \leq n-k$ and $a+k < b \leq \min(a+2k, n)$. For $n \geq 2k$ there are $(k)((n-k)-k) + (k)(k-1)/2$ such terms.

Case 4: $((b,a),(a,a+k))$, where $1 \leq a \leq n-k$ and $\max(1, a-k) \leq b < a$. For $n \geq 2k$ there are again $(k)((n-k)-k) + (k)(k-1)/2$ such terms.

Cases 3 and 4 can coincide for $n \geq 2k$ when $|a-b| = k$. All other combinations of cases are disjoint. There are precisely $(n-k)-k$ such coincidences. This gives the following recurrence for $Q_{k,1}$:

$$Q_{k,1} = 2(n-k)(k-1) + (k-1)(k) + (2k-1)(n-2k) + Q_{k-1,1}$$
$$Q_{1,1} = n-2.$$

The recurrence can be solved by standard techniques resulting in

$$Q_{k,1} = 2k^2 n - \frac{5}{3}k^3 - kn + \frac{2}{3}k - k^2.$$

Note that $|Q_k| = \binom{|P_k|}{2}$. Since Q_k is the disjoint union of $Q_{k,0}$ and $Q_{k,1}$, then

$$|Q_{k,0}| = |Q_k| - |Q_{k,1}|.$$

The variance of Φ_w can then be written as

$$\begin{aligned}\mathrm{Var}[\Phi_w] &= \frac{1}{|P_k|^2}\left(\sum_{(a,b)\in P_k}\mathrm{Var}[X_{a,b}] + 2\sum_{((a,b),(c,d))\in Q_{k,0}}\mathrm{Cov}[X_{a,b}X_{c,d}] + 2\sum_{((a,b),(c,d))\in Q_{k,1}}\mathrm{Cov}[X_{a,b}X_{c,d}]\right)\\ &= \frac{1}{|P_k|^2}(|P_k|\mathrm{Var}[X_{a,b}] + 2|Q_{k,0}|\mathrm{Cov}[X_{a,b}X_{c,d}] + 2|Q_{k,1}|\mathrm{Cov}[X_{a,b}X_{a,c}]).\end{aligned}$$

Noting that $\mathrm{Cov}[X_{a,b}X_{c,d}] = E[X_{a,b}X_{c,d}] - E[X_{a,b}]E[X_{c,d}]$ and $\mathrm{Var}[X_{a,b}] = E[X_{a,b}^2] - E[X_{a,b}]^2$, the constants c_1, c_2, and c_3 can be solved for using the relations from the previous lemma. Since the quantities u, v, and w can be computed in $O(n^2)$ time, so can the variance and expectation. ■

APPENDIX B

The rate of recombination is here referred to as $\rho = 4Nrt$, where r is the per base recombination rate and t is the sequence length. Here N was set to 1000 (diploid population), t was set to 1000 as well, and r solved for accordingly.

For population growth $\rho^\dagger$ was obtained so that the expected number of recombinations was equal under scenarios (*i.e.*, $E_{\beta=5000}[R(m)] = E_{\beta=0}[R(m)]$), where $R(m)$ is the number of recombinations for a sample of size m (Wiuf *et al.* 2001), and $\beta = Nb$, where b is the population growth rate per generation (Wiuf *et al.* 2001). The expected number of recombinations for $\beta = 0$ can be found by the following formula (Hudson and Kaplan 1985):

$$E_{\beta=0}[R(m)] = \rho\sum_{j=1}^{m-1}\frac{1}{j}.$$

Table B1 shows the values used for $\rho = 1$ (when $\beta = 0$). For values of $\rho > 1$ (*e.g.*, $\rho = 2$) one can simply double the values in the table.

Similarly, the rate of mutation is here referred to as $\theta = 4N\mu t$, where μ is the per base mutation rate and t is the sequence length. Under a Jukes–Cantor model if $\beta = 0$ then

$$\theta = t\frac{3p}{3-4p}$$

(Wiuf *et al.* 2001). This allows θ to be found for a fixed amount of sequence diversity p. For $\beta = 5000$ the appropriate value of θ was found by simulation. The values used are shown in Table B2.

TABLE B1

Conversion of the rate of recombination ρ between β = 0 and β = 5000

		ρ	
Sample size	$E[R(m)]$	β = 0	β = 5000
$m = 5$	2.08	1	550
$m = 10$	2.83	1	400
$m = 15$	3.25	1	325
$m = 25$	3.78	1	250
$m = 50$	4.48	1	175

TABLE B2

Conversion of the rate of mutation θ between β = 0 and β = 5000

	θ	
Diversity (%)	β = 0	β = 5000
$p = 1$	10.1	6,600
$p = 5$	53.6	33,000
$p = 10$	115.4	68,000
$p = 15$	187.5	106,000
$p = 25$	375	193,600